Spécialité Agronomie

Spécialisation Management des Systèmes d'Information

Mémoire de fin d'études

Formation Ingénieur AgroSup Dijon

Formation Initiale

Calcul d'indicateurs à partir de la BDNI et pistes d'optimisation du système d'information de l'US-ODR

Establishment of indicators with the BDNI and study about the optimization of the information system of the US-ODR

(Stage réalisé du 1er avril au 30 septembre)

Fanny LASSAGNE

3ème année

Ludovic JOURNAUX

Maître de mémoire

Didier RABOISSON

Maître de stage

Ecole Nationale Vétérinaire de Toulouse

23 chemin des Capelles – BP 87614

F – 31076 Toulouse Cedex 3

2013

Préface

Pour l'accueil que j'ai reçu à l'INRA je remercie toute l'équipe du service de l'observatoire du développement rural de l'institut national de recherche agronomique de Toulouse. Merci aussi à l'Ecole Nationale Vétérinaire de Toulouse pour m'avoir proposé ce stage qui m'a permis d'apprendre énormément au sujet des bases de données.

Pour l'aide technique qu'ils m'ont apportée au cours de mon stage, pour leur disponibilité pour répondre à mes questions et leurs conseils sur l'optimisation de mes requêtes SQL ainsi que pour leur patience, merci à Mme Maigné, M. Chartier, M. Garcia, M. Gendre, M. Picinbono et à mon maître de stage M. Raboisson. Cette aide m'a permis d'accélérer considérablement mon travail.

Pour les informations et les conseils qu'ils m'ont donnés concernant la Business intelligence et la mise en place d'une telle solution informatique, je remercie Mme Raynal, M.Bimonte et M.Decloix.

Pour leurs conseils et leur aide pour la rédaction de ce mémoire, merci à Mme Toulon, M. Journaux, M. Simon et M.Steffe.

Pour leur patience et l'aide qu'ils m'ont apportée avec les multiples relectures de mon mémoire ainsi que pour leurs conseils, je remercie ma famille et mes amis.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

A

ACID (propriétés) : Atomicité, cohérence, isolation, durabilité

B

BDD : Base de données

BDNI : Base de données nationale d'identification

BI : Business intelligence (informatique décisionnelle)

C

CIRAD : Centre de coopération international en recherche agronomique pour le développement

CNRS : Centre national de recherche scientifique

D

DOLAP : Desktop on-line analytical processing

DOS (commande) : Disk operating system

DW : Data warehouse (entrepôt de données)

E

ENVT : Ecole nationale vétérinaire de Toulouse

ETL : Export-transform-load

F

FTP : File transfert protocole (Protocole de transfert de fichiers)

H

HOLAP : Hybrid on-line analytical processing

I

INRA : Institut national de recherche agronomique

INSERM : Institut national de la santé et de la recherche médicale

M

MOLAP : Multidimensional on-line analytical processing

N

NOSQL : Not only structured query language

O

ODR : Observatoire du développement rural

OLAP : On-line analytical processing

OLTP : On-line transactional processing

P

PHP : Hypertext preprocessor

R

ROLAP: Relational on-line analytical processing

S

SAE2 : Sciences sociales, agriculture & alimentation, espace & environnement

SGBD : Système de gestion de bases de données

SGBDR : Système de gestion de bases de données relationnel

SI : Système d'information

SOLAP : Spatial on-line analytical processing

SQL : Structured query language

U

UMR : Unité mixte de recherche

US-ODR : Unité de service « Observatoire du développement rural »

Glossaire

A

ACID (propriétés) : Les propriétés ACID (Atomicité, cohérence, isolation, durabilité) sont des règles qui doivent être respectées par les systèmes de gestion de bases de données traditionnels.

B

BDD : Ensemble structuré permettant de stocker et manipuler (insérer, supprimer, modifier, rechercher) une grande quantité d'informations.

BDNI : Base de données nationale d'identification bovine. Base de données contenant les informations sur les bovins présents sur le territoire français. Ces informations concernent les bovins, leurs propriétaires, les exploitations où ils se trouvent, leur parenté, leurs mouvements... La BDNI de l'INRA provient de la BDNI du Ministère de l'Agriculture et est utilisée presque uniquement pour la réalisation de statistiques et la mise en place d'indicateurs et non pas pour la traçabilité des bovins au niveau français. Par la suite, le terme BDNI fera référence à celle de l'INRA, sauf mention contraire.

BI : La Business intelligence (informatique décisionnelle) est un ensemble de moyens, d'outils et de méthodes permettant de regrouper, manipuler et analyser de très grandes masses de données pour avoir une vue d'ensemble de ces informations.

D

Dictionnaire de données : Il référence l'ensemble des données d'une base de données et leurs caractéristiques, comme le nom des tables, le nom des colonnes et leur type, ainsi que des exemples d'enregistrements.

DOLAP : Le Desktop on-line analytical processing est une analyse multidimensionnelle pour laquelle les données sont stockées en local.

DOS : C'est un système d'exploitation en lignes de commandes plutôt qu'avec des interfaces (comme c'est le cas pour Windows)

Doublon : Enregistrement présent en plusieurs exemplaires dans une même table.

DW : Data warehouse. Voir « entrepôt de données ».

E

Entrepôt de données : C'est une base de données permettant de stocker, normaliser et manipuler de très grandes masses de données dans le but de les analyser.

ETL : Les ETL (Export-transform-load) sont des outils permettant de récupérer, normaliser et insérer des données provenant de sources hétérogènes (base de données, fichiers textes…) dans un entrepôt de données.

F

FTP : On l'appelle file transfert protocole ou protocole de transfert de fichiers. Ce logiciel permet d'échanger des fichiers entre deux ordinateurs du même réseau.

H

HOLAP : L'HOLAP (Hybrid on-line analytical processing) est une analyse multidimensionnelle des données utilisant à la fois des méthodes de stockage relationnelle et multidimensionnelle (ROLAP et MOLAP).

Hypercube : Il est une des représentations schématique des analyses multidimensionnelles de données.

I

Index : Il permet, lorsque l'on recherche certaines données, d'éviter de parcourir toutes les informations d'une table (du premier enregistrement au dernier) mais de procéder par dichotomie ce qui accélère de beaucoup l'exécution des requêtes.

Indicateur : C'est un nombre évalué en fonction d'une échelle de mesure de performances. Il permet d'étudier une situation dans un contexte donné.

Informatique décisionnelle : Voir « BI ».

Interface : Elle permet à un utilisateur humain de pouvoir échanger facilement des informations avec une machine.

M

Machine virtuelle : Elle simule l'existence d'un ordinateur dans un autre ordinateur. Elle permet, entre autres, de pouvoir utiliser un système d'exploitation différent de celui de l'ordinateur « hôte » et donc d'utiliser des logiciels incompatibles avec ce dernier. Elle permet aussi de réaliser des tests sans risquer des problèmes sur la machine physique, c'est-à-dire l'ordinateur.

Modèle en étoile : C'est un modèle d'organisation des tables d'une base de données décisionnelle où toutes les tables de dimensions sont reliées directement à la table de faits.

Modèle en flocon : C'est un modèle d'organisation des tables d'une base de données décisionnelle où la table de faits est reliée à plusieurs tables de dimensions, elles-mêmes reliées à d'autres tables de dimensions. Ainsi, les tables de dimension sont hiérarchisées entre elles.

MOLAP : Le MOLAP (Multidimensional on-line analytical processing) est une analyse multidimensionnelle qui stocke directement les données dans un format multidimensionnel.

N

NOSQL : Le NOSQL est un type particulier de systèmes de gestion de bases de données. Ce système permet de gérer des bases de données géantes mais ne respecte pas les règles ACID et le langage de manipulation des données n'est généralement pas le SQL.

O

ODR : Observatoire du développement rural. Ensemble d'outils informatiques permettant le regroupement et l'analyse de données sur le développement rural.

OLAP : L'OLAP est une analyse multidimensionnelle des données.

OLTP : L'OLTP permet de modifier des informations.

P

PHP : C'est un langage permettant de générer des pages web au contenu variable (par exemple en français ou en anglais selon notre choix) et/ou d'envoyer des requêtes SQL.

Plateforme décisionnelle : Une plateforme décisionnelle permet un accès centralisé et sécurisé à des données via une interface web

R

Reporting : Les outils de reporting (ou restitution) permettent de mettre en forme des résultats, par exemple sous forme de rapports présentant des graphiques et des courbes.

ROLAP : Le ROLAP (Relational on-line analytical processing) est une analyse multidimensionnelle qui stocke les données dans un format relationnel.

S

SAE2 : Département des Sciences Sociales, Agriculture & Alimentation, Espace & Environnement. Un des 14 départements de l'INRA. Son principal objectif est l'étude du monde économique et social et des politiques agricoles.

SGBD : Logiciel permettant de manipuler les informations de plusieurs bases de données.

SGBDR : C'est un système de gestion de bases de données dont les tables d'une même base sont reliées entre elles par un ou plusieurs enregistrements.

SI : UN SI (Système d'information) est un ensemble de ressource permettant de collecter, organiser, traiter et diffuser des informations.

SOLAP : Le SOLAP (Spatial on-line analytical processing) permet une analyse multidimensionnelle des données suivant un axe spatial via un affichage cartographique.

SQL : C'est un langage permettant d'envoyer des requêtes à un SGBD.

Système d'exploitation : C'est un ensemble de programmes qui démarre dès l'allumage de l'ordinateur et qui permet de l'utiliser facilement. Il en existe un grand nombre, comme Windows, Linux ou Mac OS.

T

Table de dimension : Une table de dimension contient des données en fonction desquelles on étudiera d'autres données d'une base décisionnelle. C'est un axe d'étude.

Table de faits : Une table de faits contient les données à analyser d'une base décisionnelle.

Tableau de bord : Le tableau de bord est un type particulier de restitution des données : tout doit tenir sur une feuille A4 ou sur un écran d'ordinateur (sous forme de graphiques, de courbes…).

U

UMR : Unité mixte de recherche. Unité de recherche regroupant des chercheurs de différents organismes de recherche travaillant en collaboration.

US-ODR : Unité de service « Observatoire du développement rural ». L'US-ODR appartient au département SAE2 de l'INRA

Introduction

Aujourd'hui en France, nous produisons suffisamment de nourriture pour nourrir toute la population. Cependant, la qualité et la traçabilité de cette dernière restent des enjeux d'actualité, comme nous avons pu le voir avec l'exemple de la présence de viande de cheval dans des produits certifiés 100% pur bœuf au début de l'année 2013. Depuis plusieurs années, la problématique se tourne donc plus vers la qualité que vers la quantité.

Les enjeux actuels de l'agronomie et de l'agroalimentaire sont donc plus particulièrement une meilleure traçabilité et une meilleure qualité nutritive, que ce soit des produits transformés (plats préparés, produits laitiers), des végétaux, ou des produits animaux (viandes, poissons). L'amélioration de la qualité des produits animaux est permise notamment par une amélioration de la santé des animaux et par les épidémiologies. De plus, le suivi des mouvements des animaux (achat, vente) permet d'établir une traçabilité de ces derniers.

Pour permettre l'amélioration de la traçabilité et de la qualité des produits alimentaires, il est important de recueillir de nombreuses données, notamment des données issues du monde agricole. Il faut suivre chaque produit fini, de sa production (exploitation ayant produit le végétal ou ayant vu naître le bovin, la provenance des graines ou encore l'origine des parents du bovin) à sa consommation (magasin où le lot a été vendu).

L'INRA (Institut national de recherche agronomique), et plus particulièrement l'US-ODR (Unité de service de l'Observatoire du développement rural), mène des études statistiques sur le monde agricole. Une partie de ces études est menée, en collaboration avec l'ENVT (Ecole nationale vétérinaire de Toulouse), sur les bovins du territoire français. Chaque année, de nombreuses informations sont ainsi récupérées et analysées notamment à partir de la BDNI (Base de données nationale d'identification bovine) pour étudier l'évolution sanitaire du cheptel bovin français et ses variations territoriales.

Il s'agit donc de s'assurer que ces informations soient à jour puis de les interroger par requêtes en vue de la production d'indicateurs. Ce sont les deux objectifs qui m'ont été assignés dans le cadre de ce stage. Ceci a également été l'occasion de réfléchir plus en profondeur sur le système d'information utilisé pour ce calcul d'indicateurs et ainsi identifier des pistes d'optimisation.

Ainsi, l'objectif de mon stage était de mettre à jour la base nationale d'identification bovine de l'INRA et de calculer des indicateurs permettant d'étudier le cheptel bovin français.

Tout d'abord, nous étudierons le contexte dans lequel s'inscrit ce projet ainsi que ses différents objectifs, puis nous analyserons plus précisément les besoins et difficultés qu'il implique, avant de voir les réponses qui peuvent être apportées pour produire les indicateurs de façon optimisée.

I- Contexte et objectifs

Pour comprendre une problématique et les enjeux d'un projet, il est important de bien cerner le contexte dans lequel il s'inscrit. Cette première partie a donc pour but de présenter les organismes d'accueil dans lesquels j'ai effectué mon étude ainsi que la demande qui m'a été formulée. J'expliquerai ensuite plus en détail ce qu'est un indicateur et son utilité afin de bien appréhender la finalité de cette étude. Enfin, pour comprendre sur quelles bases repose actuellement la production des indicateurs, le système d'information en place sera présenté.

I-1- Présentation des organismes d'accueil

Au cours de ce stage, j'ai travaillé sous la tutelle de l'ENVT (Ecole nationale vétérinaire de Toulouse), représentée par M. Didier Raboisson, mon maître de stage. Cependant mon poste était situé sur le site de l'INRA (Institut national de recherche agronomique) de Toulouse, plus particulièrement au sein de l'US-ODR (Observatoire du développement rural).

I-1-a- L'ENVT

L'Ecole nationale vétérinaire de Toulouse est en premier lieu un établissement de soin animalier et d'enseignement supérieur. Par ailleurs, c'est également un établissement de recherche. Elle emploie environ 75 enseignants chercheurs.

L'ENVT complète les recherches scientifiques menées sur son campus grâce à son partenariat avec de grands organismes de recherche, comme l'INRA, l'INSERM ou le CNRS. Elle est organisée en unités propres de recherche ainsi qu'en UMR (Unités mixtes de recherche) lorsqu'elle bénéficie du concours d'autres organismes [Ecole Nationale Vétérinaire de Toulouse].

L'ENVT est un des membres fondateurs du consortium pour l'agriculture, l'alimentation, la santé animale et l'environnement, « Agreenium », créé en 2009. Ce consortium a pour but de fédérer la formation et la recherche agronomique et vétérinaire au niveau national. L'ENVT coopère avec l'INRA dans ce consortium, ainsi qu'avec le CIRAD et des établissements d'enseignement supérieur agronomiques [Ecole Nationale Vétérinaire de Toulouse].

Les recherches scientifiques menées à l'ENVT contribuent à la production de nouvelles connaissances dans les domaines de la santé publique et de la santé animale. Ces connaissances concernent les enjeux liés, entre autres, aux maladies infectieuses ou génétiques, à la sécurité des aliments et à leur rôle sur la santé publique … [Ecole Nationale Vétérinaire de Toulouse].

I-1-b- L'INRA

En 1946, après la seconde guerre mondiale, « nourrir la France » est devenue une priorité. L'INRA a alors été fondé pour répondre à cette demande. Aujourd'hui, c'est le premier institut de recherche agronomique en Europe et le deuxième dans le monde. L'INRA mène des recherches au service d'enjeux de société majeurs : l'alimentation, l'agriculture et l'environnement [Institut National de Recherche Agronomique].

L'INRA est un établissement public placé sous la tutelle conjointe du ministère chargé de l'Agriculture, et des ministères chargés de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche [Service public].

Ses principales missions sont [Institut National de Recherche Agronomique] :

- de produire et diffuser des connaissances scientifiques
- de concevoir des innovations et des savoir-faire pour la société
- d'éclairer, par son expertise, les décisions des acteurs publics et privés
- de développer la culture scientifique et technique et participer au débat science/société
- de former à la recherche et par la recherche

I-1-c- Le département SAE2

Le département SAE2 (Sciences Sociales, Agriculture & Alimentation, Espace & Environnement) est l'un des quatorze départements de l'INRA. Il est majoritairement composé d'économistes et de sociologues.

Ses trois principaux objectifs sont les suivants :

- la description et la compréhension du monde économique et social
- l'explicitation des décisions des acteurs privés et publics par l'élaboration et la mise en œuvre d'instruments conceptuels et opérationnels (comme des indicateurs)
- l'explicitation des débats sur les politiques publiques et les négociations européennes et internationales en rapport avec le monde rural

Ils sont répartis en cinq domaines de recherche [Annuaire des Laboratoires et des Recherches] :

- la consommation, la transformation et la distribution des produits agroalimentaires
- les productions et les marchés agricoles
- les ressources naturelles et l'environnement
- le développement des territoires ruraux
- les approches économiques et sociales des risques, des innovations et du développement durable

I-1-d- L'unité de service US-ODR

L'US-ODR (Unité de service « Observatoire du développement rural ») appartient au département SAE2 (Sciences sociales, agriculture & alimentation, espace & environnement) de l'INRA. Elle a été créée et mise en service sous forme d'essais le premier janvier 2009, pour une durée initiale de quatre ans puis elle a été définitivement adoptée au terme de ce délai. Son principal objectif est de développer, maintenir et gérer une plateforme de données, appelée ODR (Observatoire du développement rural) [O.D.R. INRA Plateforme Recherche]. Comme son nom l'indique, cet observatoire regroupe des informations se référant aux différentes politiques agricoles et environnementales ainsi qu'aux activités agricoles et au développement rural en général.

L'US-ODR développe plusieurs observatoires sur la même plateforme, avec différents partenaires (conventionnés ou non, tiers agréés…). Ces observatoires mettent le serveur de données à disposition des équipes de recherche s'intéressant à la politique agricole et au développement rural (selon les conventions signées avec les fournisseurs de données) [O.D.R. INRA Plateforme Recherche].

Pour ce faire, les observatoires permettent aux chercheurs d'accéder aux BDD (Bases de données) et d'utiliser les informations de celles-ci pour calculer des indicateurs afin d'étudier l'évolution du développement rural. C'est notamment le cas de l'observatoire du développement rural à partir duquel les chercheurs peuvent accéder à la BDNI (Base de données nationale d'identification bovine) de l'INRA.

C'est donc sous la tutelle de l'ENVT que j'ai effectué mon stage, mais celui-ci s'est entièrement déroulé au service ODR, dans le département SAE2 de l'INRA de Toulouse. Ces deux organismes travaillent en collaboration pour recueillir et stocker de nombreuses données dans le but de calculer des indicateurs sur le développement rural.

A présent que les organismes d'accueil dans lesquels j'ai effectué mon travail ont été présentés, je vais aborder ici le but précis de cette étude.

I-2- But de l'étude : le calcul d'indicateurs

Au travers de son unité « Observatoire du développement rural » l'INRA de Toulouse regroupe des données sur le monde agricole comme la localisation et la raison sociale des exploitations, leurs activités, le nombre d'animaux d'élevage de chaque espèce qu'elles possèdent. La BDNI de l'INRA est une copie de celle du Ministère de l'Agriculture. Elle est mise à jour chaque année avec des données reçues du Ministère. Les chercheurs effectuent ensuite des statistiques et calculent des indicateurs. Ceux-ci permettent de faire un état des lieux d'une situation dans un contexte particulier et, au besoin, de mettre en place des plans d'actions visant à améliorer la situation.

La demande qui m'a été formulée était de mettre à jour la BDNI de l'INRA et de corriger les incohérences des données pour ensuite catégoriser les données et calculer des indicateurs en fonction de ces catégories. Ces indicateurs doivent être calculés par année, pour que les chercheurs puissent suivre leur évolution dans le temps. Or l'actuelle BDD ne permet pas ces calculs en fonction des années. En effet, les données concernant chaque année sont toutes regroupées ensemble dans une seule table (par exemple la table bovins contient des bovins nés de 1980 à 2011). Cependant, au vu du nombre d'enregistrements et de la lenteur d'exécution des requêtes, il est impossible d'en faire aboutir une ayant plusieurs clauses à la fois (sélection des données correspondant à une année en particulier puis reste de la requête).

L'US-ODR a souhaité que mon travail s'effectue dans la continuité de l'architecture existante. J'ai donc travaillé sur la BDD déjà présente. Cependant, pour atteindre les objectifs qui m'avaient été assignés, j'ai dû faire face à quatre difficultés :

- Des données manquantes
- Des données incohérentes
- Des problèmes technologiques
- De nouveaux indicateurs à calculer par année

En plus de la demande initiale, j'ai donc jugé utile d'étudier d'autres solutions techniques.

Ainsi l'US-ODR et l'ENVT souhaitent étudier l'évolution du monde agricole, et à travers la BDNI, l'évolution sanitaire du cheptel bovin français et ses variations territoriales, à l'aide d'indicateurs. Pour pouvoir bien appréhender les différents objectifs de cette étude, il est important de bien comprendre ce qu'est un indicateur.

I-3- Qu'est-ce qu'un indicateur et à quoi sert-il ?

I-3-a- Définition d'un indicateur

Un indicateur est un nombre, une série de nombres ou un ensemble de données permettant d'étudier une situation et/ou son évolution [Larousse]. C'est aussi l'association d'un critère d'appréciation et d'une échelle de mesure d'une performance [Guy Y., 2007]. Il apporte des informations permettant d'apprécier une situation dans un contexte particulier et de prendre des décisions collectives pour apporter éventuellement une correction à cette situation.

Ainsi, les données de la BDNI sont utilisées pour calculer différents indicateurs comme des taux de mortalité par race, en fonction du sexe… ou comme des nombres moyens de bovins par atelier et par exploitation.

Les indicateurs de l'INRA interviennent pour la réalisation de cartes géographiques descriptives (Figure 1). Celles-ci permettent de comparer géographiquement des résultats obtenus dans un même contexte et de les rendre accessibles à la fois aux chercheurs et aux statisticiens mais aussi aux agriculteurs, médias et par ce biais, à la plupart des gens puisque représentés sous forme de gradients de couleur sur une carte géographique. En réalisant des cartes thématiques sur différentes années on peut aussi voir l'évolution des résultats de cet indicateur au cours du temps. L'INRA utilise les indicateurs pour faire un état des lieux d'une situation dans un contexte particulier ainsi que pour calculer des facteurs de risque et pour éventuellement mettre en place des plans d'actions visant à améliorer la situation.

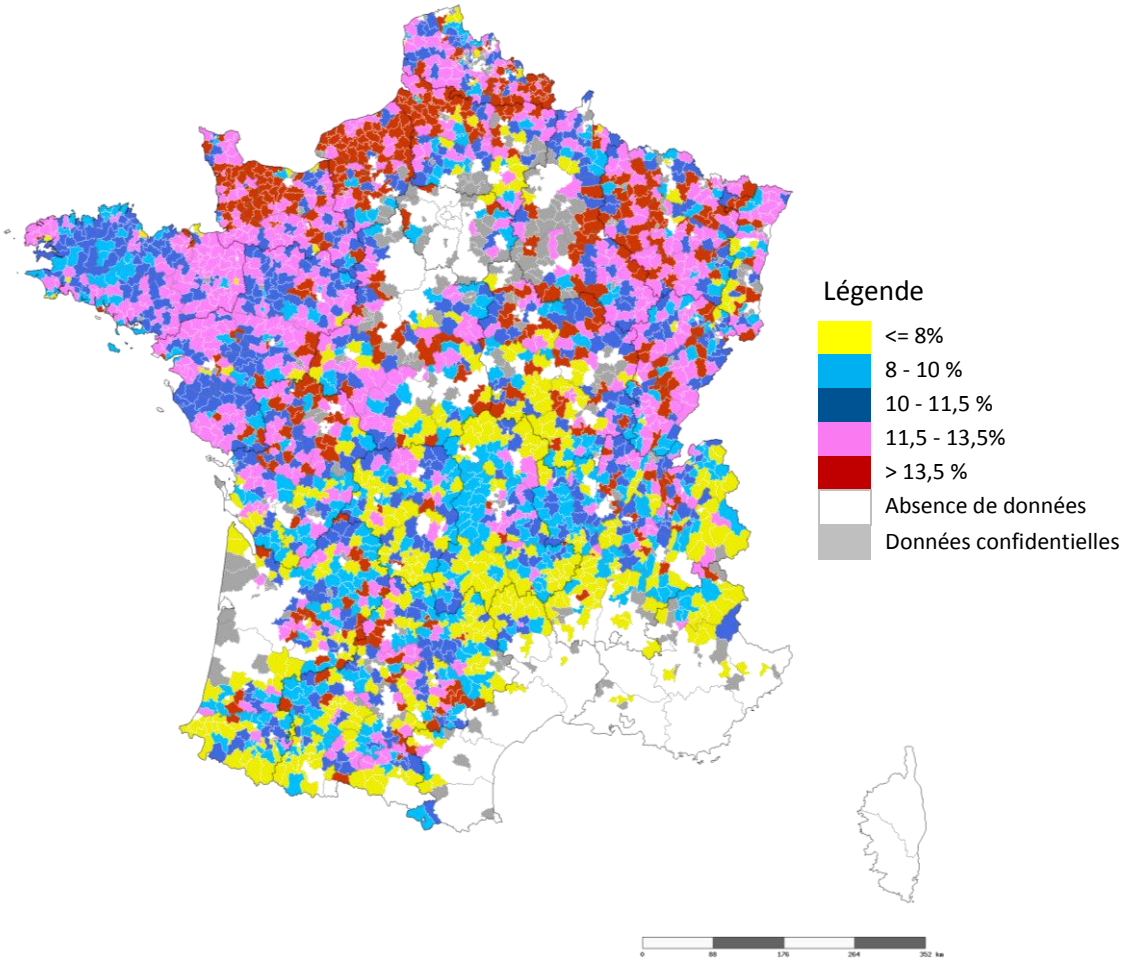


Figure 1: taux de mortalité des veaux de race laitière par canton français en 2011

Mon rôle sera de calculer des indicateurs permettant d'étudier l'évolution sanitaire du cheptel bovin français et ses variations territoriales à partir des données de la BDNI. Un de ces indicateurs est le taux de mortalité.

I-3-b- Exemple d'indicateur : le taux de mortalité

Cet indicateur est fréquemment calculé, notamment pour étudier l'évolution sanitaire du cheptel bovin français et ses variations territoriales.

Il est calculé en divisant le nombre d'animaux morts dans une exploitation sur une période de temps donné par le nombre total de bovins détenus par cette exploitation sur cette même période. Cet indicateur peut être calculé sur des périodes de temps plus ou moins longues (un mois, un semestre, une année) et les résultats peuvent être groupés par zone géographique selon différentes échelles (exploitation, canton, département, région).

Les zones géographiques ayant un taux de mortalité élevé peuvent alors prendre des décisions collectives pour mettre en place des solutions pour améliorer la gestion de leur cheptel et diminuer ce taux de mortalité.

Un indicateur est donc un nombre évalué en fonction d'une échelle de mesure de performances. Il permet d'étudier une situation dans un contexte donné et peut être mis sous forme de cartes géographiques statistiques pour être compris par les chercheurs et statisticiens ainsi que par les agriculteurs et par n'importe quelle personne s'y intéressant. A l'INRA, les indicateurs sont calculés à partir d'informations stockées dans des bases de données.

I-4- Etude du SI de l'INRA et de la BDNI, les supports du calcul d'indicateurs

L'objectif de cette partie est de comprendre la structure de la BDNI et du système d'information ainsi que les enjeux et attentes des chercheurs l'utilisant. Pour ce faire, j'ai interrogé des personnes ayant des compétences et connaissances différentes les unes des autres afin d'avoir un aperçu le plus complet possible du matériel et de ce qui était utilisé à l'INRA.

I-4-a- Le système d'information

L'INRA utilise majoritairement des logiciels/applications/serveurs gratuits et Open Source (dont la distribution et la création de produits dérivés est libre mais pas forcément gratuite).

Serveur Linux Debian :

Une machine physique (ordinateur) est située à l'INRA de Montpellier et héberge 2 machines virtuelles (simulations d'ordinateur à l'intérieur de la machine physique) sur lesquelles sont installés des serveurs sous Linux Debian (version Lenny) : esrcarto et

esrcarto2 (Figure 2). Chaque machine virtuelle possède 8 processeurs et 32 Gio de mémoire vive ainsi qu'un espace disque de 1 To.

Lors de la création des machines virtuelles, l'INRA de Montpellier a choisi Linux Debian car ce serveur avait la réputation d'être très stable.

Machine virtuelle esrcarto2 :

C'est la machine virtuelle sur laquelle je travaille, en parallèle avec d'autres chercheurs. Elle possède le SGBD MySQL accessible via l'interface phpMyAdmin. Avec cette machine chacun travaille sur ses propres bases de données.

Inconvénient : lorsque certaines requêtes envoyées au serveur sont trop lourdes, elles entrainent une baisse de performance du serveur qui se répercute sur l'exécution des requêtes de tous les utilisateurs se servant de cette machine virtuelle.

Serveur web Apache :

La machine virtuelle sur laquelle je travaille (esrcarto2) dispose d'un serveur web Apache (serveur Open Source et gratuit).

Interface : phpMyAdmin

Cette interface graphique est accessible sur la machine virtuelle esrcarto2 via le serveur web Apache (Figure 2). Elle permet d'accéder au SGBD MySQL et d'en manipuler les données facilement. En effet, il existe des onglets reprenant les requêtes les plus fréquemment utilisées comme la création et la suppression de BDD, de tables ainsi que de colonnes.

SGBD MySQL :

L'US-ODR utilise deux SGBD : PostGreSQL est utilisé pour son extension PostGis permettant entre autres la réalisation de cartes statistiques et MySQL est utilisé pour tout le reste. Le SGBD sur lequel j'ai travaillé sur la machine virtuelle esrcarto2 est MySQL. Il est accessible via l'interface phpMyAdmin (Figure 2) ainsi que par commande DOS.

L'INRA l'a choisi au départ car c'était un outil Open Source, connu et simple à manipuler.

Moteur de stockage : MyISAM

Le moteur de stockage de MySQL stockant les bases de données avec lesquelles j'ai travaillé est MyISAM. Lorsque la BDNI de l'INRA a été créée, c'est MyISAM qui a été utilisé (moteur par défaut de l'époque) car c'était le moteur de stockage le plus utilisé par l'INRA. Ce moteur de stockage est bien adapté à la gestion des données statistiques de l'US-ODR pour lesquelles il n'est pas nécessaire d'avoir toutes les capacités d'un SGBD classique (comme la gestion des clefs étrangères par exemple). Par ailleurs, il est très efficace pour les jointures de tables, très fréquentes pour la réalisation de statistiques et simple à manipuler. Ici, l'utilité des BDD n'est pas leur fonction relationnelle, mais bien leur possibilité de stockage de centaines de Gio de données.

Avec ce moteur de stockage, chaque table est stockée sous forme de 3 fichiers : un .frm, un .MYI et un .MYD. Sous le .frm (pour « format ») est enregistrée la structure de la table, comprenant, entre autres, son nom, ceux de ses colonnes, la longueur et le type de données de ses colonnes. Le .MYI (pour MYIndex) quant à lui stocke les index tandis que le .MYD (pour MYData) conserve les données enregistrées dans la table.

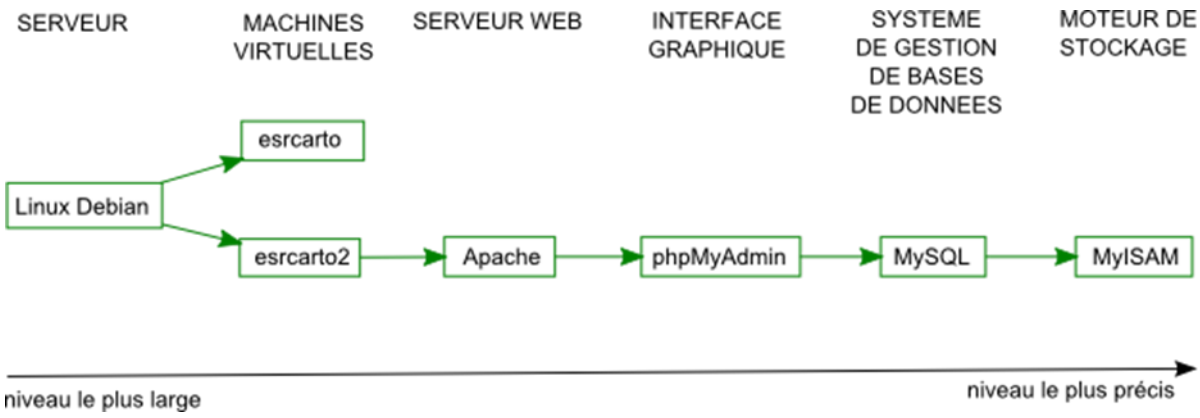


Figure 2: résumé des outils en place à l'INRA de Toulouse

Langages :

L'US-ODR utilise deux langages pour ses diverses activités : PHP (Hypertext preprocessor) et SQL (Structured query language). Pour l'interrogation et la manipulation des bases de données, c'est le SQL qui est utilisé, soit directement sur phpMyAdmin soit intégré dans des programmes PHP.

L'intérêt du langage SQL est sa normalisation : il permet de gérer différents SGBD presque de la même manière.

Gestionnaire de fichiers : Total Commander

Total Commander est un gestionnaire de fichiers (tout comme l'explorateur de Windows). Pour ma part, je ne m'en suis servie que pour décompresser certains fichiers de données et surtout pour avoir un aperçu du contenu de certains de ces fichiers, trop lourds pour être ouverts par des éditeurs de texte classiques (comme Bloc Note ou Notepad).

Après avoir déterminé les outils utilisés à l'US-ODR, j'ai cherché à comprendre la manière dont la sécurité était gérée au sein de ce service de l'INRA.

Gestion de la sécurité et des sauvegardes

Dans le service de l'ODR, les sauvegardes sont gérées de la manière suivante :

- Tous les soirs, les bases de données sont sauvegardées sur le serveur où elles sont déjà présentes
- Une fois par semaine, le vendredi, elles sont sauvegardées sur un autre serveur

Par ailleurs, la sécurité est gérée de différentes manières : seules les adresses IP autorisées peuvent accéder aux machines virtuelles, l'accès aux bâtiments se fait par un badge et pour se connecter aux machines virtuelles il faut connaître les identifiant et mot de passe, communs à tous les agents INRA. La sécurité pourrait être encore améliorée en donnant un identifiant et mot de passe unique à chaque utilisateur INRA, ce qui permettrait de suivre qui s'est connecté à la machine.

La sécurité est donc gérée de différentes manières, par adresses IP et identifiants de connexion, et deux types de sauvegardes sont réalisés très régulièrement. Une fois ces informations très importantes assimilées, je me suis intéressée à la matière à proprement parler de mon travail, c'est-à-dire aux données, et plus précisément à la structure et au contenu de la BDNI déjà présente et aux données manquantes reçues.

I-4-b- Les données

Afin de mieux comprendre la BDNI et son contenu, sur lesquels j'allais travailler, j'ai établi le dictionnaire des données de la bdni10 (Annexe 1) ainsi que celui des données reçues du ministère de l'agriculture (Annexe 2). Ils référencent l'ensemble des données de ces bases et leurs caractéristiques, ainsi que des exemples d'enregistrements.

Ces dictionnaires de données m'ont permis de mieux comprendre la structure des différentes tables ainsi que leur contenu. J'ai aussi pu établir les correspondances entre les colonnes de la bdni10 et ceux des données reçues afin de pouvoir importer ces dernières.

La base de données bdni10

Cette base de données est gérée par le SGBD MySQL de la machine virtuelle esrcarto2. L'utilité de la bdni10 (Base de données nationale d'identification version 1.0) n'est pas sa fonction relationnelle, mais sa possibilité de stockage d'une grande masse d'informations (elle fait 56,1 Gio). Ces données sont ensuite analysées et utilisées pour la mise en place d'indicateurs permettant le suivi du développement rural.

A mon arrivée à l'US-ODR il existait un dictionnaire de données de la bdni10 référençant l'ensemble de ses données et leurs caractéristiques, comme le nom des tables, le nom des colonnes et leur type. Cependant il était incomplet et présentait plusieurs différences avec la structure réelle de la base. Afin de mieux comprendre à quoi correspondait chacune des colonnes, j'ai comparé la bdni10 à ce dictionnaire des données et j'ai mis ce dernier à jour. A partir de celui-ci j'ai pu réaliser le diagramme des classes correspondant (Figure 3).

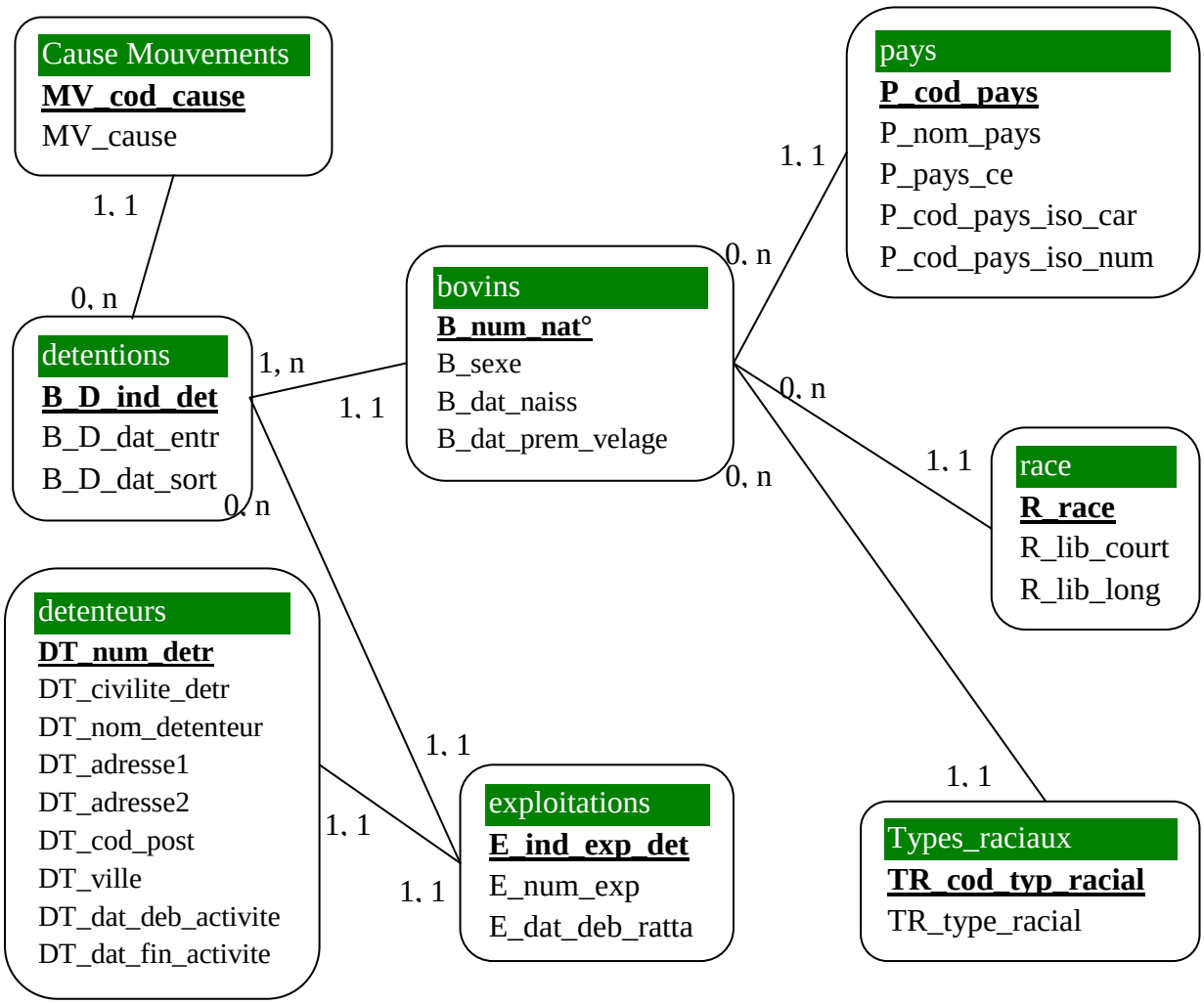


Figure 3: diagramme des classes de la bdni10

Certaines tables de la bdni10 comprenaient plusieurs colonnes servant d'identifiants. Avec l'accord de mon maître de stage, il a été décidé de n'en garder qu'une seule. Par exemple, la table bovins, en plus d'avoir la colonne B_num_nat (numéro national) unique pour chaque bovin, possédait une colonne B_ind_bovin qui était auto incrémentée et elle aussi unique pour chaque bovin. Ici j'ai juste gardé le numéro national qui suffisait à lui seul à identifier un animal. Par ailleurs des incohérences ont été repérées au sein de ces données.

Je me suis ensuite intéressée aux données de 2007 pour voir quelles colonnes pouvaient correspondre à celles de la bdni10 et être importées.

Les données de 2007

Les données de 2007 concernant les bovins et reçues du ministère de l'agriculture sont sous forme de fichiers textes (Figure 4). Il y a 12 fichiers dont la taille varie de 1,5 Mo à 1,5 Gio pour un poids total de 5,5 Gio environ.

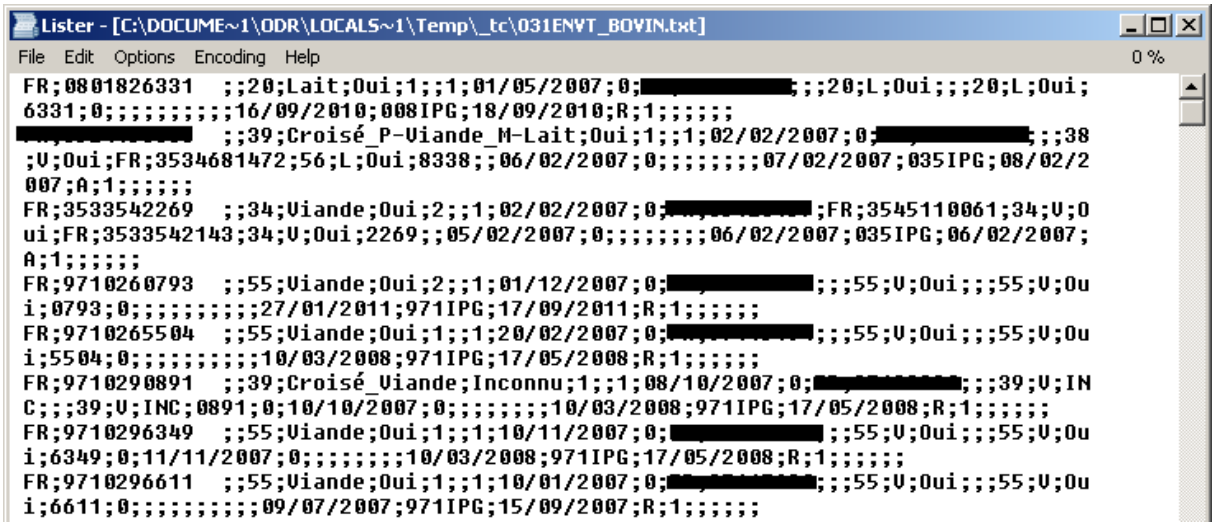


Figure 4: aperçu du contenu d'un fichier texte source des données de 2007 (les numéros d'exploitations ont été cachés)

Afin de voir quelles colonnes pouvaient être importées, j'ai créé le dictionnaire des données des fichiers reçus. Avec ces fichiers, un document Excel nous avait aussi été envoyé. Il correspondait à un dictionnaire des données partiel et contenait les noms de certaines tables et de leurs colonnes, leur type de donnée et leur longueur. Avec Excel j'y ai rajouté un exemple pour chaque colonne de la table ainsi qu'une explication claire sur ce à quoi elle correspondait car les noms n'étaient pas tous explicites.

Pour voir les données des fichiers textes (pour en faire le dictionnaire des données) qui étaient trop gros pour être ouverts avec des logiciels classiques, sur le conseil de plusieurs chercheurs j'ai utilisé le logiciel Total Commander, déjà installé sur les machines.

Ainsi la bdni10 de l'US-ODR est une BDD contenant des informations sur les bovins français jusqu'à l'année 2011 mais ayant des données manquantes pour 2007. Les trois tables clés de cette base sont les tables bovins, détentions et exploitations (nous verrons pourquoi par la suite). Les informations reçues concernant l'année 2007 sont, quant à elles, sous la forme de plusieurs fichiers textes.

L'US-ODR et l'ENVT souhaitent étudier l'évolution du monde agricole (et plus particulièrement l'évolution sanitaire du cheptel bovin français et ses variations territoriales) avec l'aide d'indicateurs. Il faut pour cela corriger et mettre à jour les données de leur base nationale d'identification bovine (BDNI) avant de catégoriser ces informations pour mettre en place ces indicateurs. Le but de ce stage est donc la mise en place pérenne d'un système de calcul d'indicateurs avec une multi-catégorisation des données. J'ai réalisé ce travail dans la continuité de l'architecture existante, comme demandé par l'INRA. Cependant, j'ai jugé utile d'étudier d'autres solutions techniques.

II- Etude des différents besoins liés à ce stage

II-1- Caractérisation du besoin

La BDNI est très principalement utilisée par mon maître de stage, M. Didier Raboisson, pour la réalisation de statistiques et de cartes géographiques descriptives. C'est lui qui m'a expliqué les besoins liés à ce stage et qui a validé tous mes choix par la suite. D'autres personnes, en général des vétérinaires ou des chercheurs, utilisent parfois la BDNI pour faire des cartographies d'indicateurs. C'est par mail que j'ai communiqué avec ces personnes.

Concrètement, les besoins de l'US-ODR par rapport à cette base sont de pouvoir :

- Insérer de nouvelles données (en général une fois par an)
- Modifier les données existantes
- Calculer des indicateurs (initialement les taux de mortalités et les nombres de bovins par atelier, puis d'autres indicateurs en fonction des besoins par la suite)

Il faut que ces actions puissent être exécutées assez rapidement (l'exécution d'une requête ne doit pas prendre plus d'une heure ou deux), malgré le grand nombre d'informations.

Afin de bien comprendre les fonctionnalités demandées par l'US-ODR, nous allons voir quelques exemples de cas d'utilisation qui décrivent des séries d'actions aboutissant à un des objectifs de cette étude.

II-1-a- Exemple de cas d'utilisation n°1 : l'insertion de données

Ce premier cas d'utilisation présente l'exemple d'un stagiaire souhaitant insérer de nouvelles données dans la BDNI. J'ai d'abord décrit le cas d'utilisation (Tableau 1) puis énoncé les règles de gestion (Tableau 2) avant de décrire les enchaînements de la réalisation de ce cas (Tableau 3).

Tableau 1: description du cas d'utilisation n°1

Cas d'utilisation	Insertion de données dans la table bovins
Acteurs ….. principaux	Stagiaire
But	Pouvoir insérer de nouvelles données reçues dans la table bovins
Résumé métier	Le stagiaire doit pouvoir facilement et rapidement insérer un ensemble de nouvelles données reçues concernant les bovins dans la table bovins déjà existante
Accès	Dans la BDNI, dans la table bovins
Pré-condition	Le stagiaire doit avoir accès à la base Les nouvelles données doivent être elles aussi sous forme de BDD.
Post-condition	
Commentaires	

Règles de gestion

Tableau 2: règles de gestion du cas d'utilisation n°1

N° Règle	Définition
1	Seuls les nouveaux bovins (bovins dont le numéro national est absent de la BDNI) pourront être insérés dans la table bovins
2	La BDNI et la base contenant les nouvelles données doivent avoir la même structure

Enchaînements

Tableau 3: enchaînements du cas d'utilisation n°1

N° Enchaînement	Scénario alternatif	Action
1		L'utilisateur accède au SGBD où se trouve la BDNI
2		L'utilisateur lance l'union de la BDNI avec la BDD contenant les nouvelles données pour créer une nouvelle BDNI contenant les nouvelles données. La requête peut se faire en lançant directement le code SQL ou bien via une interface (type phpMyAdmin)

II-1-b- Exemple de cas d'utilisation n°2 : la mise à jour de données

Ce deuxième cas d'utilisation présente l'exemple d'utilisateurs (stagiaire, chercheurs…) souhaitant mettre à jour des données dans la BDNI (l'éventuelle correction d'incohérences se déroulera de manière à peu près similaire). J'ai d'abord décrit le cas d'utilisation (Tableau 4) puis énoncé les règles de gestion (Tableau 5) avant de décrire les enchaînements de la réalisation de ce cas (Tableau 6).

Tableau 4: description du cas d'utilisation n°2

Cas d'utilisation	**Mise à jour des données sur les bovins**
Acteurs **….. principaux** **….. *secondaires***	Stagiaire *Chercheurs*
But	Pouvoir modifier les données de la table bovins dans la BDNI pour les mettre à jour
Résumé métier	Le stagiaire ou les chercheurs doivent pouvoir modifier les données de la base lorsque celles-ci doivent être mises à jour ou lorsque des incohérences ont été repérées.
Accès	Dans la BDNI, dans la table bovins
Pré-condition	L'utilisateur doit avoir accès à la base. Les données contenant les mises à jour doivent être sous forme de BDD.
Post-condition	
Commentaires	

Règles de gestion

Tableau 5: règles de gestion du cas d'utilisation n°2

N° Règle	**Définition**
1	La BDNI et la base contenant les données avec les mises à jour doivent avoir la même structure

Enchaînements

Tableau 6: enchaînements du cas d'utilisation n°2

N° Enchaînement	**Scénario alternatif**	**Action**
1		L'utilisateur accède au SGBD où se trouve la BDNI
2		L'utilisateur lance la modification des données de la BDNI avec les données de la base qui contient les mises à jour. La requête peut se faire en lançant directement le code SQL ou bien via une interface (type phpMyAdmin)

II-1-c- Exemple de cas d'utilisation n°3 : création d'un atelier « veau de boucherie »

Ce cas d'utilisation présente l'exemple d'un stagiaire souhaitant préciser la présence ou l'absence de différents ateliers (« veaux sous la mère », « engraissement »…) dans les exploitations. Ici est présenté l'exemple de l'atelier « veau de boucherie ». J'ai d'abord décrit le cas d'utilisation (Tableau 7) puis énoncé les règles de gestion (Tableau 8) avant de décrire les enchaînements de la réalisation de ce cas (Tableau 9).

Tableau 7: description du cas d'utilisation n°3

Cas d'utilisation	Attribution de l'atelier « veau de boucherie » en 2007
Acteurs ….. principaux	Stagiaire
But	Pouvoir valider la présence d'un atelier « veau de boucherie » en 2007 dans certaines exploitations
Résumé métier	Le stagiaire doit pouvoir valider la présence d'un atelier « veau de boucherie » dans les exploitations agricoles lorsque celles-ci respectent les conditions nécessaires
Accès	Dans la BDNI, dans la table exploitations
Pré-condition	Les données sont cohérentes. L'utilisateur doit avoir accès à la base. Les conditions d'attribution de l'atelier ont été définies préalablement
Post-condition	Chaque exploitation possède la colonne « VB » (pour « veau de boucherie ») renseignée (resp. 1 ou 0) selon si elle possède (resp. ne possède pas) au moins 6 veaux qualifiés « de boucherie » en 2007.
Commentaires	

Règles de gestion

Tableau 8: règles de gestion du cas d'utilisation n°3

N° Règle	Définition
1	La colonne « VB » se retrouvera dans chaque ligne exploitation mais possèdera 2 valeurs en fonction de la présence ou de l'absence de l'atelier en 2007.
2	La valeur par défaut est 0 (on considère que l'exploitation ne possède pas l'atelier).
3	La valeur 1 est attribuée lorsque l'exploitation respecte les conditions d'obtention définies.

Enchaînements

Tableau 9: enchaînements du cas d'utilisation n°3

N° Enchaînement	Scénario alternatif	Action
1		L'utilisateur accède au SGBD où se trouve la BDNI
2		L'utilisateur crée une nouvelle colonne « VB » dans la table exploitations avec 0 comme valeur par défaut
3		L'utilisateur identifie (par requête) les veaux qualifiés « de boucherie » détenus en 2007
4		L'utilisateur identifie (par requête) les exploitations ayant au moins 6 veaux de boucherie en 2007
5		L'utilisateur change la valeur de la colonne « VB » de ces exploitations en 1

II-1-d- Exemple de cas d'utilisation n°4 : calcul d'un indicateur

Ce cas d'utilisation présente l'exemple d'utilisateurs souhaitant calculer un indicateur. L'exemple choisi est le taux de mortalité des veaux de boucherie, présents dans des exploitations ayant l'atelier « veaux de boucherie », et détenus en 2007. J'ai d'abord décrit le cas d'utilisation (Tableau 10) puis énoncé les règles de gestion (Tableau 11) avant de décrire les enchaînements de la réalisation de ce cas (Tableau 12).

Tableau 10 : description du cas d'utilisation n°4

Cas d'utilisation	Calcul du taux de mortalité des veaux de boucherie en 2007
Acteurs ….. principaux	Chercheurs, statisticiens
But	Pouvoir calculer le taux de mortalité des veaux de boucherie détenus au cours de l'année 2007
Résumé métier	Les chercheurs et statisticiens doivent pouvoir calculer le taux de mortalité des veaux de boucherie (et présents dans les ateliers « veaux de boucherie ») détenus au cours de l'année 2007
Accès	Dans la BDNI, dans la table detentions
Pré-condition	L'utilisateur doit avoir accès à la base. Les données cibles (veaux de boucherie appartenant à des exploitations ayant un atelier « veau de boucherie ») sont présélectionnées ou catégorisées (on connait les exploitations ayant l'atelier « veaux de boucherie » ainsi que les veaux de boucherie détenus par ces exploitations en 2007)
Post-condition	
Commentaires	

Règles de gestion

Tableau 11: règles de gestion du cas d'utilisation n°4

N° Règle	Définition
1	Le calcul du taux de mortalité ne doit se faire que sur les veaux détenus au moins 1 jour en 2007.
2	Le calcul du taux de mortalité ne doit se faire que sur les veaux de boucherie des exploitations ayant un atelier « veaux de boucherie »

Enchaînements

Tableau 12: enchaînements du cas d'utilisation n°4

N° Enchaînement	Scénario alternatif	Action
1		L'utilisateur accède au SGBD où se trouve la BDNI
2		L'utilisateur identifie les veaux de boucherie détenus en 2007 par des exploitations ayant un atelier « VB » cette année-là
3		L'utilisateur identifie les veaux morts « accidentellement » (pas envoyés à la boucherie ni consommés par l'exploitant) parmi les veaux cités précédemment
4		L'utilisateur calcule le taux de mortalité

Ainsi, les besoins exprimés par rapport à ce stage sont principalement de pouvoir insérer de nouvelles données dans la BDNI, modifier les données existantes et calculer des indicateurs. Il faut que ces actions puissent être exécutées assez rapidement (moins de deux heures pour l'exécution d'une requête) et que le calcul d'indicateurs soit fiable. Les données à partir desquelles les indicateurs sont calculés doivent donc être à jour et cohérentes.

II-2- Les traitements qui en découlent

Avant de calculer les indicateurs, des traitements des données sont nécessaires. En effet, des incohérences ont été repérées dans les données, et la BDNI doit être mise à jour avec les nouvelles données reçues du Ministère de l'Agriculture.

II-2-a- Les incohérences des données

Présence d'une date de 1er vêlage pour les mâles

Pour l'ensemble des bovins mâles j'ai cherché à détecter si une date de premier vêlage avait été renseignée. En effet, un bovin dont le sexe renseigné est « mâle », ne devrait pas avoir de date de premier vêlage renseignée. Or cette erreur est apparue pour 4 579 bovins au total.

Cette incohérence peut avoir deux sources : ou bien le bovin est en réalité une femelle et la date de vêlage a bien lieu d'être, ou bien c'est un mâle et la date de vêlage ne devrait pas exister. Une partie de ces bovins avaient bien au moins un veau enregistré dans la BDNI de l'INRA, mais d'autres n'en avaient pas.

Sexe non renseigné

Pour l'ensemble des bovins j'ai cherché à détecter si certains n' avaient pas de sexe renseigné (c'est-à-dire que dans la colonne « sexe » l'enregistrement est « null » au lieu d'être « mâle » ou « femelle », ce qui pose problème pour la classification des données demandée par la suite par l'INRA).

C'était le cas pour 297 d'entre eux (sur l'ensemble des années). Après vérification j'ai pu noter qu'aucun de ces bovins n'avait une date de premier vêlage renseignée ni aucun veau enregistré dans la BDNI.

Bovins qui ont vêlé avant de naître

Pour l'ensemble des bovins j'ai cherché à détecter si certains avaient une date de premier vêlage antérieure à leur date de naissance. Dans ce cas, c'est qu'au moins une des deux dates a été mal renseignée.

C'était le cas de 34 bovins sur l'ensemble des bovins de bovins_bdni12. Pour dix de ces vaches il existe au moins un veau présent dans notre BDNI donc pour celles-ci on peut comparer la date de leur premier vêlage avec la date de naissance de leur premier veau (sous réserve que la date de naissance des veaux ait été bien renseignée). Mais pour les autres vaches, rien n'est sûr.

Bovins qui ont une date de premier vêlage avant l'âge de 18 mois

Pour l'ensemble des bovins j'ai cherché à détecter si certains avaient une date de premier vêlage avant l'âge d'un an et demi (18 mois) car dans les faits, il est rare que ce soit le cas. Lorsque cela arrive, il est possible que ce soit dû à un mauvais renseignement de la date de naissance et/ou de premier vêlage mais il se peut aussi que les dates soient correctes. C'était le cas de 38 861 bovins au total.

Bovins qui ont une date de premier vêlage avant l'âge d'un an

Pour l'ensemble des bovins j'ai cherché à détecter si certains avaient une date de premier vêlage avant l'âge d'un an car il est très peu probable que ce soit réellement le cas. Comme précédemment, il est possible que ce soit dû à un mauvais renseignement de la date de naissance et/ou de premier vêlage mais il se peut aussi que les dates soient correctes. Ceci concernait 102 bovins au total

Détentions en double avec exploitations différentes

Il s'agit de voir s'il existe des détentions concernant le même animal et les mêmes dates d'entrée et sortie mais des exploitations différentes. Dans ce cas-là, une des deux détentions est en trop en plus d'avoir mal été renseignée.

Par exemple pour 2007, ça a été le cas de 6 détentions (au total 3 détentions concernant chacune un bovin différent, mais existant chacune en double avec des numéros d'exploitations renseignés différents). Les résultats par année sont visibles dans le tableau 13.

Tableau 13: détentions en double avec exploitations différentes

	J2004	2005	2006	2007	2008	2009	2010	2011
détentions	0	0	8	6	14	8	8	4

Date d'entrée antérieure à la date de naissance

Il s'agit, pour chaque table detent_yyyy, de voir s'il existe des bovins détenus avant d'être nés. Ceci concernait environ 50 détentions par table detent_yyyy. L'erreur est due au fait que l'une des deux dates (voire même les deux dates) a été mal renseignée.

Détentions intérieures

Il s'agit, pour chaque table detent_yyyy, de voir si pour un même bovin il existe des détentions d2 à l'intérieur d'une autre détention d1 (Figure 5). Par la suite nous considérerons que la première détention est celle ayant la plus ancienne date d'entrée et la deuxième détention est celle ayant la date d'entrée la plus récente. Sur la figure suivante, DE correspond à la date d'entrée, DS à la date de sortie et les chiffres indiquent s'il s'agit de la première ou deuxième détention. L'incohérence provient du fait qu'une date (ou plusieurs) a été mal renseignée.

Figure 5: représentation d'un chevauchement de détentions dites "intérieures"

Les nombres de détentions intérieures pour chaque table detent_yyyy est décrit dans le Tableau 2.

Tableau 14: nombre de détentions intérieures

		d1 (détentions contenant les d2)							
		j2004	2005	2006	2007	2008	2009	2010	2011
d2 (détentions contenues)	j2004	88	90	89	82	57	59	60	47
	2005	26	99	82	246	0	50	94	91
	2006	19	28	27	307	2	74	144	170
	2007	40	83	114	201	96	26	11	5
	2008	4	8	16	154	11	88	176	211
	2009	5	0	0	24	0	24	0	0
	2010	1	0	0	19	0	0	9	0
	2011	1	0	0	5	0	0	0	15

Chevauchements avec la première date de sortie nulle

Il s'agit, pour chaque table detent_yyyy, de voir si pour un même bovin il existe un chevauchement entre des détentions dû à l'absence de date de sortie dans la première détention (Figure 6).

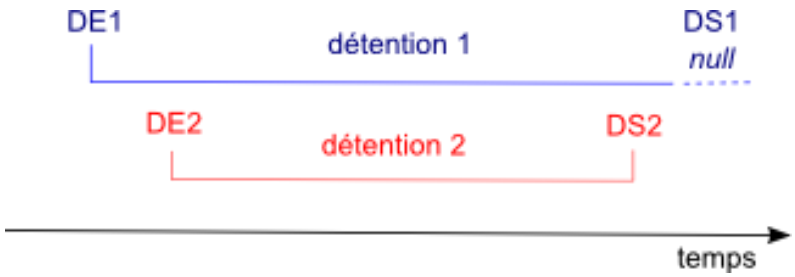


Figure 6: représentation d'un chevauchement de détentions dont la date de sortie de la première détention est nulle

Dans ce cas on suppose que la date de sortie de sortie de la première détention n'a pas été renseignée par oubli. Les différents nombres par année sont dans le tableau 15.

Tableau 15: nombre de chevauchement de détentions dont une date de sortie est nulle

	J2004	2005	2006	2007	2008	2009	2010	2011
détentions	3 195	7 401	12 797	13 038	13 918	14 582	13 171	11 628

Chevauchements de détentions

Il s'agit, pour chaque table detent_yyyy, de voir si pour un même bovin il existe un chevauchement entre des détentions (Figure 7) : le bovin apparaît dans une première détention ayant une date d'entrée et une date de sortie renseignées puis dans une deuxième, qui commence à l'intérieur de la première détention et qui se finit après (ici la deuxième date de sortie peut être nulle). L'incohérence provient du fait qu'une date (ou plusieurs) a été mal renseignée.

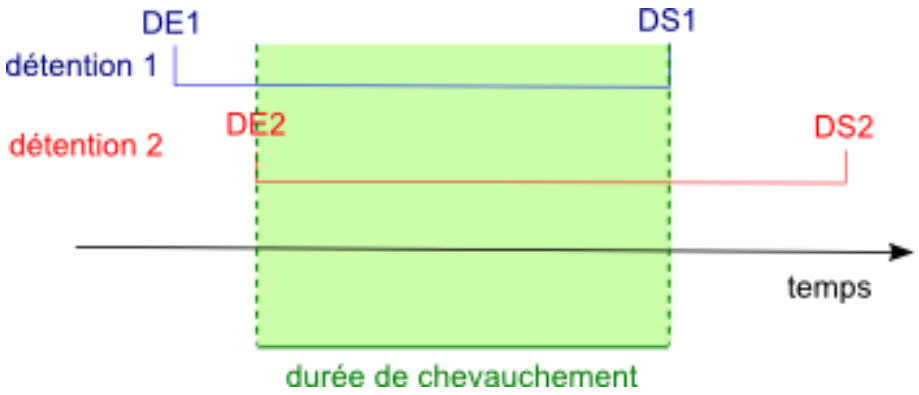


Figure 7: situations différentes en fonction de la position de la date de premier vêlage par rapport au chevauchement des détentions

Pour l'ensemble des bovins, le nombre de chevauchements par année est le suivant (Tableau 16) :

Tableau 16: nombre de chevauchements de détentions par année

		d2 (commence pendant d1 et se termine après sa fin)						
		2005	2006	2007	2008	2009	2010	2011
d1 (commence avant d2 et se termine pendant d2)	2005	11 845	5 766	37 765	10 709	8 558	7 474	5 929
	2006	17	22 271	56 522	12 291	10 990	9 958	8 289
	2007	6 151	14 606	67 604	76 569	59 365	51 714	46 412
	2008	22	61	14 699	59 403	39 668	27 402	21 523
	2009	11	24	11 017	181	56 661	46 325	31 535
	2010	5	23	5 772	16	221	55 810	49 229
	2011	1	20	2 442	7	30	243	52 683

Mort unique par bovin

Il s'agit de voir si pour un bovin il existe plusieurs détentions se terminant par sa mort, qu'elle soit naturelle ou non.

Sur l'ensemble des détentions pour les bovins nés en 2007 et après, les nombres d'erreurs étaient les suivants : 238 944 bovins avec 2 morts, 225 avec 3 morts, 1 avec 4 morts.

L'erreur peut provenir d'un mauvais renseignement de la cause de sortie (par exemple on met « mort » au lieu de « vente »).

Doublons

Il s'agit de voir dans chacune des tables detent_yyyy s'il existe des détentions en tous points identiques : même bovin, même exploitation, mêmes dates et causes d'entrée et de

sortie (Tableau 17). Ces erreurs peuvent être dues au fait qu'une personne ait entré plusieurs fois le même enregistrement ou bien qu'il a été dédoublé lors de manipulations des données.

Tableau 17: nombre de doublons par année

	2005	2006	2007	2008	2009	2010	2011
Doublons	3 575	10 511	10 524	3 608	9	9	5

Estimation du taux d'erreurs par rapport aux données correctes

J'ai cherché à estimer la proportion d'erreurs incohérentes au sein des données (en particulier au sein des tables bovins et detentions). J'ai donc fait une estimation du nombre d'incohérences repérées sur ces tables (par année). Puis j'ai sommé le nombre d'enregistrements des tables bovins et detentions par année et j'ai calculé le taux d'incohérences dans les données des tables bovins et detentions (Tableau 18).

Il est important de noter que certaines incohérences ont pu m'échapper et que ces nombres ne sont que des estimations

Tableau 18: estimation du taux d'incohérences dans les données des tables bovins et detentions

Année	2005	2006	2007	2008	2009	2010	2011
Nombre d'incohérences repérées estimé (en milliers)	97	129	149	129	116	114	108
Nombre total de détentions + bovins nés (en milliers)	39 062	38 920	38 545	38 781	38 212	38 173	37 789
Taux d'erreurs (en %)	0,25	0,33	0,39	0,33	0,30	0,30	0,28

Le pourcentage de données incohérentes sur l'ensemble des données des tables bovins et detentions est donc en moyenne de 0,32%.

Cependant, bien que ce nombre soit assez faible, ces données ne peuvent rester dans cet état. Deux possibilités apparaissent alors : la correction de celles-ci ou bien leur suppression. Le problème de la correction des données est que même si celles-ci deviennent plus cohérentes, elles n'en sont pas plus justes (puisque la correction se fait en suivant un plan de correction et non en recherchant la valeur correcte de l'enregistrement). De son côté, la suppression impactera un nombre de données supérieur à 0,32% : en effet, prenons l'exemple d'un bovin dont on a repéré un chevauchement entre deux de ses détentions. En supposant que ce bovin ait d'autres détentions avant et après celles-ci, ce sera l'ensemble de toutes ces détentions qu'il faudra supprimer car sinon en suivant la vie du bovin on pourra voir ses premières détentions puis on aura un « trou » au milieu de sa vie et enfin d'autres détentions et cela risque d'impacter les résultats par la suite.

Ainsi, des données incohérentes existent dans la BDNI et si les indicateurs sont calculés à partir de ces dernières, ils ne seront pas fiables. Avant de calculer les indicateurs, il faut donc soit corriger les données, ce qui les rend plus logiques mais pas plus fiables, soit les supprimer. Cependant, en plus de supprimer les données incohérentes, il faut aussi supprimer les données qui leur sont liées pour ne pas avoir de « trous » dans les informations. En plus de ces traitements, la BDNI est mise à jour annuellement.

II-2-b- L'alimentation de la base

Début 2013, l'US-ODR a reçu de nouvelles données concernant les bovins en 2012. Ces données sont à intégrer à la BDD déjà en place. Le principal problème pour cette incorporation est la taille importante des tables : certaines possèdent près de 100 millions d'enregistrements.

Par ailleurs, une partie des données de 2007 étant absente de la base, il faut aussi la mettre en forme et l'importer dans la BDNI. Cependant, la bdni10 contient déjà une partie des données de 2007 et il ne faudrait pas que ces dernières soient présentes en plusieurs exemplaires. Il faudra donc veiller à ce que seuls les enregistrements absents soient importés. Il faudra aussi faire attention à ce que des enregistrements à mettre à jour ne soient pas considérés comme totalement nouveaux (du fait de quelques différences).

Pour ces données concernant l'année 2007, il faudra créer une base spécifique et les importer. Ensuite, la réalisation d'un dictionnaire des données permettra de voir les correspondances avec les colonnes de la bdni10.

Dans un deuxième temps, l'importation des données non présentes dans la bdni10 dans une nouvelle BDD contenant déjà ces anciennes données pourra se faire, ainsi que les mises à jour et d'éventuelles modifications.

Import des données de 2007

Après avoir étudié la structure des données de 2007 avec un dictionnaire des données (Annexe 2), je les ai importées localement dans des tables créées par mes soins sur EasyPHP, une plateforme de développement web avec laquelle on peut faire fonctionner des scripts PHP en local. EasyPHP possède phpMyAdmin et MySQL. Pour créer ces tables, j'ai rédigé un script PHP que j'ai placé dans le répertoire C:\easyphp\mysql\data\bdni_2007. Ensuite, en ouvrant l'interface phpMyAdmin, il suffisait de cliquer sur « bdni_2007 » puis sur le script de mon choix pour le lancer (Figure 8).

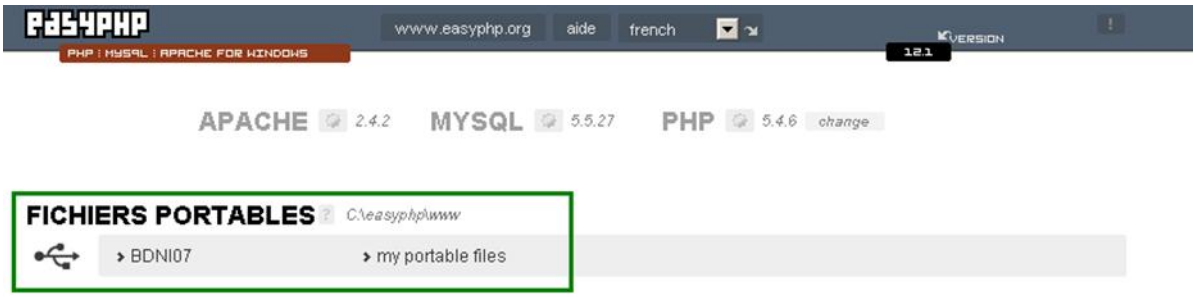


Figure 8: lancer un script PHP dans EasyPHP

Les différentes tables ont ainsi été créées très rapidement. Avec les tables dans MySQL j'ai pu évaluer la place nécessaire dans esrcarto2 et voir si elle était suffisante. Les données de 2007, une fois mises sous forme de BDD, avaient un poids de 12,5 Gio.

Pour la correction et le traitement des données j'ai choisi de découper les tables par année. En effet, la réalisation des indicateurs se faisant par année, il fallait obtenir, à la fin des traitements, des tables divisées par année, en partant initialement de très grosses tables regroupant les données sur toutes les années. Deux choix apparaissent alors : faire d'abord les imports, corrections et traitements et ensuite diviser les tables volumineuses par année ou bien diviser dès le départ les tables volumineuses pour faire ensuite les imports, corrections et traitements. L'avantage de la première méthode est quelle limitait le nombre de lignes de code (et donc le temps passé à l'écrire) mais que l'ordinateur traitait les requêtes beaucoup plus lentement que sur de petites tables contenant nettement moins d'enregistrements (pour rappel, les requêtes se faisaient sur des tables dont le nombre de lignes variait en général de 1 million à 100 millions). La deuxième méthode, quant à elle, accélérait très sensiblement la vitesse de traitement des requêtes par l'ordinateur mais faisait se multiplier le nombre de tables et de lignes de code par sept (les années allant de 2005 à 2011).

Après avoir essayé quelques requêtes sur les tables d'environ 100 millions d'enregistrements je me suis aperçu que le nombre d'enregistrements était tel que la requête n'aboutissait pas (parfois elle tournait encore, deux jours après l'avoir lancée). J'ai donc choisi la deuxième méthode, celle qui consistait à d'abord diviser les tables pour ensuite faire les imports, corrections et traitements.

Division de la table detentions en detent_yyyy

Par la suite, les noms de tables se terminant par « _yyyy » concernent les tables pour l'ensemble des années. Par exemple, au lieu de parler de detent_2005, detent_2006, … et detent_2011, je mentionnerai uniquement detent_yyyy ce qui limitera les répétitions.

Les tables detent_yyyy contiennent les détentions des bovins. Une détention concerne un seul bovin (reconnu par son numéro national) avec une date et une cause d'entrée (comme la naissance ou l'achat) dans une exploitation (reconnue par son numéro Insee). Elle peut éventuellement contenir une date et une cause de sortie (comme la mort ou la vente).

Différentes modalités

Pour expliquer ces différentes modalités d'importation, je vais prendre l'exemple de la table detent_2007. Dans bdni12.detent_2007 (créée à partir de bdni10.detentions) j'ai importé les détentions dont au moins 1 jour se trouve en 2007. Pour cela, j'ai déterminé trois modalités : il fallait que la date d'entrée en détention (DE) soit antérieure à 2007 et que sa date de sortie (DS) soit en 2007 ou inexistante ou alors que ses dates d'entrée et de sortie soient toutes les deux en 2007. J'ai procédé de la manière suivante (Tableau 19) :

Tableau 19: modalités d'importation des détentions dans les tables detent_yyyy

Date d'entrée	Date de sortie	Inclus dans la table ?
<= 31.12.2006	<= 31.12.2006	Non
<= 31.12.2006	>= 01.01.2007	Oui (Mod 1)
<= 31.12.2006	= NULL	Oui (Mod 2)
>= 01.01.2007 et <= 31.12.2007		Oui (Mod 3)
> 31.12.2007		Non

Pour les années antérieures à 2005

Dans la bdni12, à partir des données de la bdni10, j'ai aussi créé une table detent_j2004 (« jusqu'à 2004 inclus ») qui contient les détentions se déroulant sur au moins 1 jour avant 2005 (même principe que pour detent_2007 expliqué précédemment mais ici la table detent_j2004 contient les détentions pour toutes les années, jusqu'à 2004 inclus.
Pour cela les modalités étaient différentes : il fallait uniquement DE <= 31.12.2004.

Division de la table bovins en bovinsnes_yyyy

Dans la bdni12, à partir des données de la bdni10 j'ai créé en parallèle :

- Une table avec tous les bovins (équivalent de bdni10.bovins) nommée bovins_tous
- Les tables bovinsnes_yyyy

Les tables bovinsnes_yyyy contiennent uniquement les bovins nés l'année yyyy correspondante. Par exemple, la table bovinsnes_2007 contient uniquement les bovins nés en 2007 (Figure 9). De son côté la table bovinsnes_j2004 contient tous les bovins nés avant 2005.

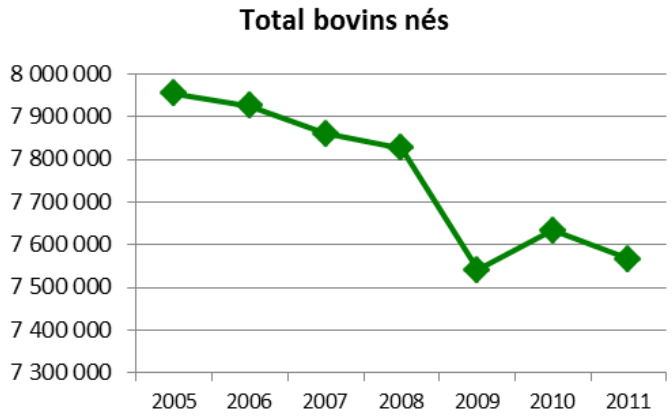


Figure 9: nombre de bovins nés par année

Division de la table exploitations en exploit_yyyy

Dans les tables exploit_yyyy de bdni12, j'ai importé les exploitations de la table bdni10.exploitations détenant au moins 1 bovin l'année yyyy (Figure 10). Ce sont donc les exploitations qui existent dans bdni12.detent_yyyy. Par exemple, dans la table exploit_2007 de bdni12, se trouvent les exploitations ayant détenu des bovins en 2007. J'ai procédé de même pour toutes les tables, y compris j2004.

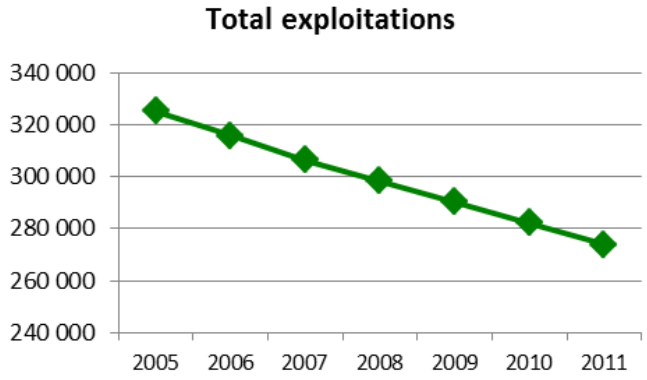


Figure 10: nombre d'exploitations par année

Afin de calculer les indicateurs par année, comme demandé par l'US-ODR, j'ai divisé les principales tables de la BDNI par année, ce qui m'a aussi permis d'accélérer le temps d'exécution des requêtes. J'ai ensuite divisé les données des tables « bovins », « detentions » et « exploitations » dans des catégories, ce qui m'a permis de calculer les indicateurs demandés. Pour ce faire, j'ai travaillé dans la continuité de l'architecture existante. Cependant, certaines difficultés sont apparues, notamment au niveau du temps d'exécution des requêtes. J'ai alors étudié d'autres solutions permettant de répondre à la problématique et pouvant se montrer plus adaptées que celle-ci.

III- Les solutions répondant aux besoins

III-1- Description des données pour trouver une solution adaptée

La BDNI contient de grandes masses de données. En effet, certaines tables possèdent près de 100 millions d'enregistrements et son poids dépasse les 56Gio. Ces données proviennent de sources différentes (BDD, fichiers textes) et sont hétérogènes : certaines sont des dates (comme la date de naissance ou la date de début de détention), d'autres sont des caractères (comme le code du pays du bovin ou le nom de l'exploitant) et d'autres encore sont des nombres (comme le sexe du bovin). Ce sont donc des informations très détaillées dont une grande partie est statique (notamment pour les tables « bovins », « exploitations » ou « exploitants ») mais quelques-unes sont dynamiques (comme les dates de premier vêlage des bovins qui sont renseignées une seule fois mais en général après l'enregistrement du bovin dans la base ou comme les dates de fin de détentions qui doivent être mises à jour). Les mises à jour sont généralement annuelles et permettent l'insertion de nouveaux enregistrements et la mise à jour d'enregistrements existants (avec actualisation des données précédentes).

Dans le SI (Système d'information) on attend de ces données qu'elles puissent être résumées et analysées et que les résultats puissent être visualisés, que ce soit sous forme de tableaux, de graphiques, de cartes géographiques. Ces résultats doivent être faciles et rapides à comprendre pour tout le monde, que l'utilisateur soit informaticien, statisticien, vétérinaire, chercheur, agriculteur, journaliste ou qu'il fasse partie du grand public. Les résultats des indicateurs doivent partir de l'échelle de l'exploitation et doivent pouvoir être agrégés en cantons, départements, régions… si nécessaire. L'agrégation doit se faire rapidement et facilement.

Peu d'utilisateurs manipulent la BDNI et ils ne sont pas dans une optique de concurrence. Ils travaillent sur des thématiques différentes.

La BDNI possède donc de très grandes masses de données provenant de sources hétérogènes qui doivent être regroupées et normalisées à des fins d'analyse et de calcul d'indicateurs. Le SI actuel permet de répondre aux besoins mais présente plusieurs difficultés, notamment au niveau du temps d'exécution des requêtes. Il est donc intéressant d'étudier différentes solutions pouvant être mises en place afin de répondre à la problématique de la manière la plus adaptée.

III-2- Choix du système d'information à mettre en place

Au regard des besoins énoncés ainsi que de la structure et de l'utilisation des données de la BDNI, j'ai étudié différentes pistes d'optimisation du système d'information environnant cette BDD. Je vais présenter ici trois solutions envisageables : les SGBD (Système de gestion de bases de données) relationnels (solution déjà en place), le NOSQL (Not only SQL) et la BI (Informatique décisionnelle).

III-2-a- Les SGBD relationnels

Un SGBD (Système de gestion de bases de données) est un logiciel permettant d'accéder aux informations d'une ou de plusieurs BDD (Bases de données) et de les manipuler. C'est un intermédiaire entre les bases de données et les utilisateurs.

Les SGBD relationnels se basent sur la répartition des données en différentes tables. Celles-ci contiennent des colonnes qui permettent de stocker les données. Les tables de la BDD sont reliées entre elles par des relations permettant d'associer un enregistrement à une ou plusieurs autres. Ce système de relation permet d'éviter les redondances de données (en utilisant les clés étrangères) ce qui limite les incohérences de données dues aux erreurs de frappe.

Une des caractéristiques des SGBD relationnels est leur respect des règles ACID (Atomicité, Cohérence, Isolation, Durabilité) [Gray J., 1981].

Atomicité : elle permet de vérifier qu'une requête est effectuée totalement ou pas du tout : si un problème apparaît lors de la requête et que celle-ci ne peut se terminer, il faut effacer toute trace de la transaction et remettre les données dans l'état où elles étaient avant le début de celle-ci. L'atomicité doit être respectée dans toutes situations, comme une panne d'électricité ou une défaillance de l'ordinateur.

Cohérence : elle permet de vérifier qu'une requête part de données cohérentes et amène à des données toujours cohérentes. Tout changement amené dans la base de données doit être valide selon toutes les règles définies.

Isolation (ou Indépendance) : Elle assure que l'exécution simultanée de plusieurs requêtes donne le même résultat que si on avait exécuté les requêtes les unes à la suite des.

Durabilité : Elle assure que lorsqu'une transaction a été confirmée, elle demeure enregistrée même à la suite d'une panne survenant immédiatement après l'exécution de la requête, que ce soit une panne d'électricité, une panne de l'ordinateur ou un autre problème.

Cependant, l'architecture doit être adaptée aux besoins des utilisateurs : en effet, avec les SGBD relationnels, plus la quantité de données est importante, plus les temps de saisie et de traitement des données seront importants (en partie car les transactions doivent se faire

dans le respect des règles ACID). Par ailleurs, lorsque l'on souhaite créer une architecture évolutive, pouvant s'adapter aux différents besoins des utilisateurs dans le futur, la BDD contient alors des données très fragmentées et donc nombreuses, ce qui allonge considérablement la durée de traitement. L'essentiel est donc d'adapter l'architecture aux besoins des utilisateurs [Gray J., 1981].

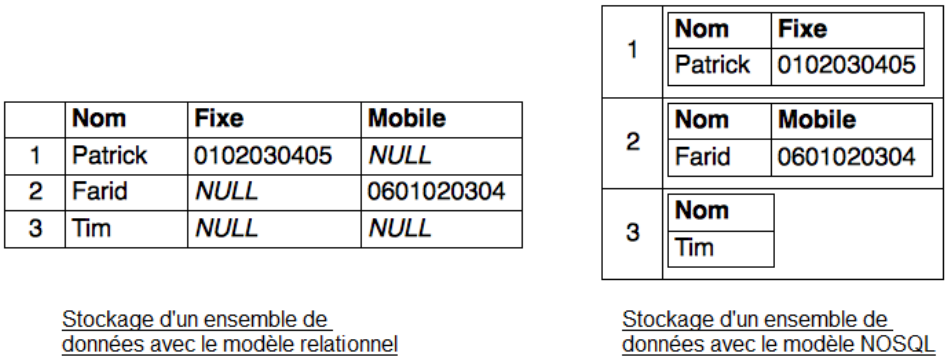
Le SGBD relationnel est le système actuellement en place (c'est MySQL qui est utilisé à l'US-ODR), mais les grands volumes de données de la BDNI posent de gros problèmes de performance.

III-2-b- Le NOSQL

Les principales différences entre les SGBD relationnels utilisant le SQL (Structured query language) et les SGBD à modèle NOSQL (Not only SQL) sont les suivantes : ces derniers ne sont pas soumis aux règles ACID et leur représentation des données n'est pas relationnelle. Ceci permet à la fois d'améliorer les performances du SGBD et d'augmenter la capacité à traiter de très grands volumes de données [TheCodingMachine]. En effet, le NOSQL est né pour faire face à des difficultés que rencontrent de plus en plus souvent les SGBD du modèle relationnel, notamment le stock d'un grand nombre de données et la gestion de données ayant une structure complexe et hiérarchisées entre elles. Pour en faire de même avec le modèle relationnel, il faudrait un ensemble de plusieurs tables relationnelles utilisant différentes clefs, et ce faisant, plus le stock de données sera important et plus les performances du SGBD se dégraderont [About.com]. Ainsi, les SGBD NOSQL sont mieux adaptés aux BDD géantes et sont d'ailleurs utilisés pour la gestion des sites web de très grande audience comme Google, Amazon, Facebook ou eBay [Downing, A., 2011].

Il existe différents types de BDD fonctionnant selon le modèle NOSQL. Parmi ces types de BDD, 4 sont principalement utilisés : les types clé-valeur, document, colonne et graphe. Je ne décrirai ici que le type colonne, qui est celui correspondant à la structure de la BDNI.

Les BDD colonne permettent d'avoir un grand nombre de valeurs sur une même ligne, ce qui permet ainsi de stocker les relations de type 1-N. Dans une base de données relationnelle, ce stockage serait sous forme de tableau, avec un nombre fixe de données (les données non renseignées auraient comme valeur *NULL*). Ici, la BDD colonne permet de stocker ces ensembles de données avec un nombre variable de données (les valeurs *NULL* ne sont pas stockées) (voir Figure 11).



	Nom	Fixe	Mobile
1	Patrick	0102030405	*NULL*
2	Farid	*NULL*	0601020304
3	Tim	*NULL*	*NULL*

Stockage d'un ensemble de données avec le modèle relationnel

1	Nom	Fixe
	Patrick	0102030405
2	Nom	Mobile
	Farid	0601020304
3	Nom	
	Tim	

Stockage d'un ensemble de données avec le modèle NOSQL

Figure 11: comparaison du stockage d'un ensemble de données avec les modèles relationnel et NOSQL [Inovia Conseil]

La plupart des développeurs sont intéressés par la manière dont une BDD NOSQL peut être interrogée, car si on ne peut sélectionner les données selon des critères parmi des milliers d'autres données, le SGBD est inutile. Un des inconvénients des SGBD NOSQL est que leur langage de requête n'est pas aussi performant que le SQL, celui des SGBD relationnel. A la place, ils proposent un système de requête spécifique au type de BDD [About.com].

Cependant, depuis peu, le SQL peu malgré être utilisé dans certains SGBD NOSQL (encore peu nombreux). C'est le cas de Hadoop, un SGBD NOSQL de type colonne [IBM].

Ainsi, le NOSQL permet de stocker les informations de la manière la plus adaptée possible et présente une meilleure performance, moins de contraintes et une montée en charge plus importante. Cependant, son langage de requête est moins performant que celui du modèle relationnel. Par ailleurs, le rôle de la BDNI étant de stocker des données pour calculer des indicateurs, un système permettant de les calculer simplement ainsi que de les mettre en forme serait particulièrement utile.

III-2-c- L'Informatique Décisionnelle

L'informatique décisionnelle, aussi appelée BI (Business intelligence), est un ensemble de moyens, d'outils et de méthodes permettant de collecter des données ainsi que de les analyser et de les restituer [Beranger F. et al., 2012]. Cette analyse des données permet de produire des indicateurs et des rapports. En général, la BI s'appuie sur un (ou des) entrepôt(s) de données (Data Warehouse) pour stocker de très grandes masses de données provenant de sources diverses (Figure 12). L'informatique décisionnelle doit donc aussi proposer des outils de navigation, d'interrogation et de visualisation de l'entrepôt [Tranchant M., 2012].

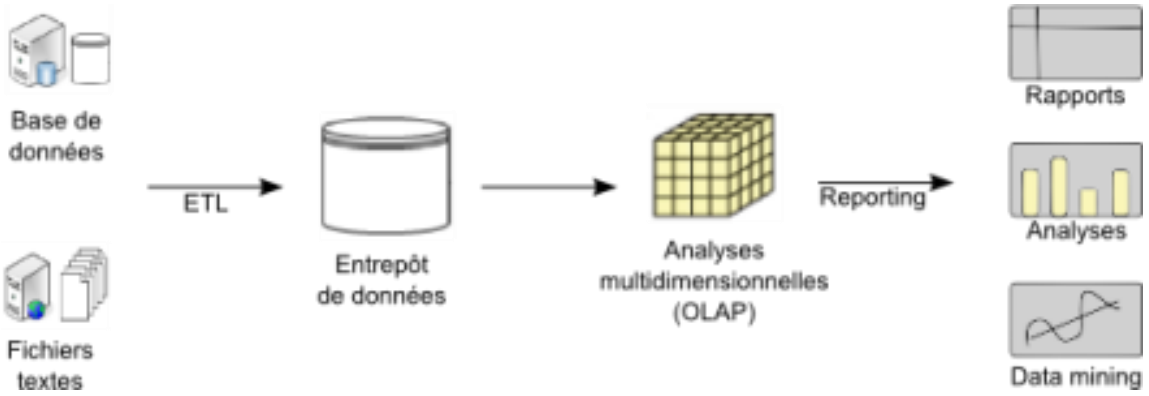


Figure 12: principes de la BI

Les sources de données sont souvent diverses et variées comme pour la BDNI de l'INRA où une partie des informations est déjà sous forme de BDD dans la BDNI et où d'autres données sont sous forme de fichiers texte. Des outils ETL (Extract-transform-load) sont ainsi nécessaires afin de les extraire, de les nettoyer, de les transformer et de les insérer dans un entrepôt de données [Taslimanka Sylla M., 2007].

Un entrepôt de données est un système de stockage des données pouvant supporter de très grandes masses d'informations. Contrairement à la plupart des bases de données classiques, il est conçu pour être interrogé à des fins d'analyse [Cyran M., Lane P. & Polk J., 2005].

Cependant, cette charge de travail pour analyser les données serait très difficilement supportable par une base de données classique (en particulier pour une BDD comme la BDNI qui possède plusieurs Gio d'informations). La solution est d'utiliser une structure plus orientée vers l'analyse, que l'on appelle un système OLAP (On-line analytical processing) [Grim Y., 2008]. L'OLAP permet de réaliser des analyses multidimensionnelles des données. Celles-ci permettent par exemple de visualiser les résultats d'un indicateur en suivant à la fois un axe temporel et un axe géographique (comme les cantons). Les analyses multidimensionnelles permettent ainsi une utilisation facile et intuitive des données pour les utilisateurs non informaticiens [Abello A. & Romero O., 2009].

Enfin, les outils de reporting permettent la restitution des résultats des analyses multidimensionnelles (généralement en quelques secondes) sous forme de rapports contenant (entre autres) des graphiques, diagrammes, courbes et tableaux [Tranchant M., 2012]. Ils permettent de visualiser facilement et rapidement ces résultats [Taslimanka Sylla M., 2007].

L'informatique décisionnelle est donc plus compliquée à mettre en place que les solutions précédentes, mais permet ensuite une intégration, une manipulation et des analyses des données beaucoup plus simples et rapides.

III-2-d- Comparaison et préconisations

Nous avons vu précédemment que le SI à mettre en place doit pouvoir supporter de grandes masses de données provenant de sources différentes (BDD et fichiers textes). Ces données seront mises à jour annuellement et devront pouvoir être corrigées facilement lorsque des incohérences seront repérées et le travail devra se faire préférentiellement en SQL. Par ailleurs, la BDNI servant à calculer des indicateurs, le SI doit permettre l'analyse de ces données et la mise en forme des résultats (sous formes de courbes, graphiques…).

J'ai donc comparé sur ces différents points (en allant de – pour une solution répondant peu ou pas au problème à +++ pour une très bonne réponse) les SGBD relationnels et le NOSQL (Tableau 20).

Tableau 20: comparaison de trois solutions pour la mise en place d'un SI efficace

	SGBD	NOSQL
Support d'une grande masse de données	+	+++
Intégration de données de sources différentes	+	+
Mise à jour et modifications faciles	++	++
Langage de requête SQL	+	-
Analyse des données	-	+
Mise en forme des résultats	-	-

La BI, quant à elle, peut reposer sur chacun de ces deux types de stockage et offre une bonne (voire une très bonne) réponse à chacun de ces points. C'est donc cette dernière que je préconise.

Ainsi, chacune des trois solutions permet de répondre à la problématique, mais la BI, qui peut s'appuyer sur les SGBD relationnels ou sur le NOSQL, permet une manipulation et des analyses des données facilitées. Cependant, sa mise en place est plus longue et nécessite de respecter certaines étapes.

III-3- Mise en place du SI préconisé

III-3-a- Modélisation de la base de données décisionnelle

Dans un système d'information classique, une BDD est composée de tables et de relations, une table étant une représentation d'une entité, d'un ensemble de données, et une relation étant un lien entre ces entités [Blain F.-A. & Grim Y., 2008]. En informatique décisionnelle, on parle en termes de tables de faits et de tables de dimensions. Pour réaliser des analyses multidimensionnelles, on utilise généralement une table de fait et plusieurs tables de dimensions [Abello A. & Romero O., 2009].

Pour déterminer quelle table est la table de faits, il faudra déterminer sur quelle table seront calculés la majorité des indicateurs [Taslimanka Sylla M., 2007]. Par exemple, pour le taux de mortalité, on fait le rapport entre le nombre de morts et le nombre total de bovins, et pour compter le nombre de morts on compte le nombre de bovins dont une détention se termine par la cause de sortie « mort ». Un fait est tout ce qu'on voudra analyser [Blain F.-A. & Grim Y., 2008].

Pour déterminer les tables de dimensions, il faut repérer quelles tables permettent le calcul des indicateurs en suivant des axes différents. Les dimensions sont les axes avec lesquels on veut faire l'analyse. Pour la BDNI ça peut être les tables exploitations ou types_raciaux par exemple. Une dimension est tout ce qu'on utilisera pour faire nos analyses [Blain F.-A. & Grim Y., 2008].

Cette division des données en tables de faits et de dimensions permet de les organiser selon une vision dimensionnelle qu'en ont les utilisateurs. Les données d'un contexte sont perçues sous la forme de matrices à deux ou plusieurs dimensions (axes temporel, géographique…). Ces matrices sont plus généralement appelées hypercubes [Gouarné J.-M., 1997]. Elles présentent plusieurs avantages, notamment une représentation intuitive des données qui les rend accessibles aux non-informaticiens, un accès plus direct aux données ce qui permet de diminuer les temps d'exécution des requêtes de l'ordinateur par rapport à un SGBDR [Gouarné J.-M., 1997].

L'organisation de ces tables de faits et de dimensions peut se faire selon plusieurs modèles dont les deux principaux sont les modèles dits « en étoile » ou « en flocon ».

Dans le modèle en étoile, toutes les dimensions sont reliées directement à la table de faits. Schématiquement, cette représention a l'aspect d'une étoile, d'où son nom [Blain F.-A. & Grim Y., 2008].

Dans le modèle en flocon, on considère qu'il peut exister des hiérarchies entre les tables de dimensions. Le table de faits est donc reliée à plusieurs tables de première

dimension dont certaines d'entre elles sont elles même reliées à des tables de deuxième dimension… [Blain F.-A. & Grim Y., 2008]

Le modèle en étoile permet de limiter le nombre de jointures des requêtes d'interrogation et sur des configurations légères, de diminuer les temps de traitements de ces requêtes [Decideo.fr]. Ceci serait un gros avantage dans le cas de la BDNI car aujourd'hui un des plus gros problèmes est la lenteur d'exécution des requêtes (et plus il y a de jointures entre tables et plus ce temps s'allonge encore).

Cependant, le modèle en flocon permet notamment de diminuer les redondances et d'intégrer une structure hiérarchique entre les données [Decideo.fr]. En limitant le nombre de redondances, on pourra bénéficier d'indicateurs plus fiables car les données seront présentes en un seul et unique exemplaire (par exemple, pour un calcul du taux de mortalité sur un canton particulier, on pourrait avoir le choix entre « nomcanton », « NomCanton » et « NOMCANTON » ce qui posera problème pour le calcul car toutes les données ne seront pas sélectionnées). En conséquence, ce modèle est le plus adapté pour la réalisation d'indicateurs par la suite. Cependant, en le choisissant, on sait que le temps d'obtention des résultats des calculs sera plus long qu'en utilisant le modèle en étoile.

Le modèle en flocon est donc la plus adaptée pour l'usage que l'on souhaite faire de la BDNI. On peut voir la représentation de la BDNI selon ce modèle dans la figure 13. La table de faits est en orange, les tables de première dimension sont envert foncé et les tables de deuxième dimension sont en vert clair.

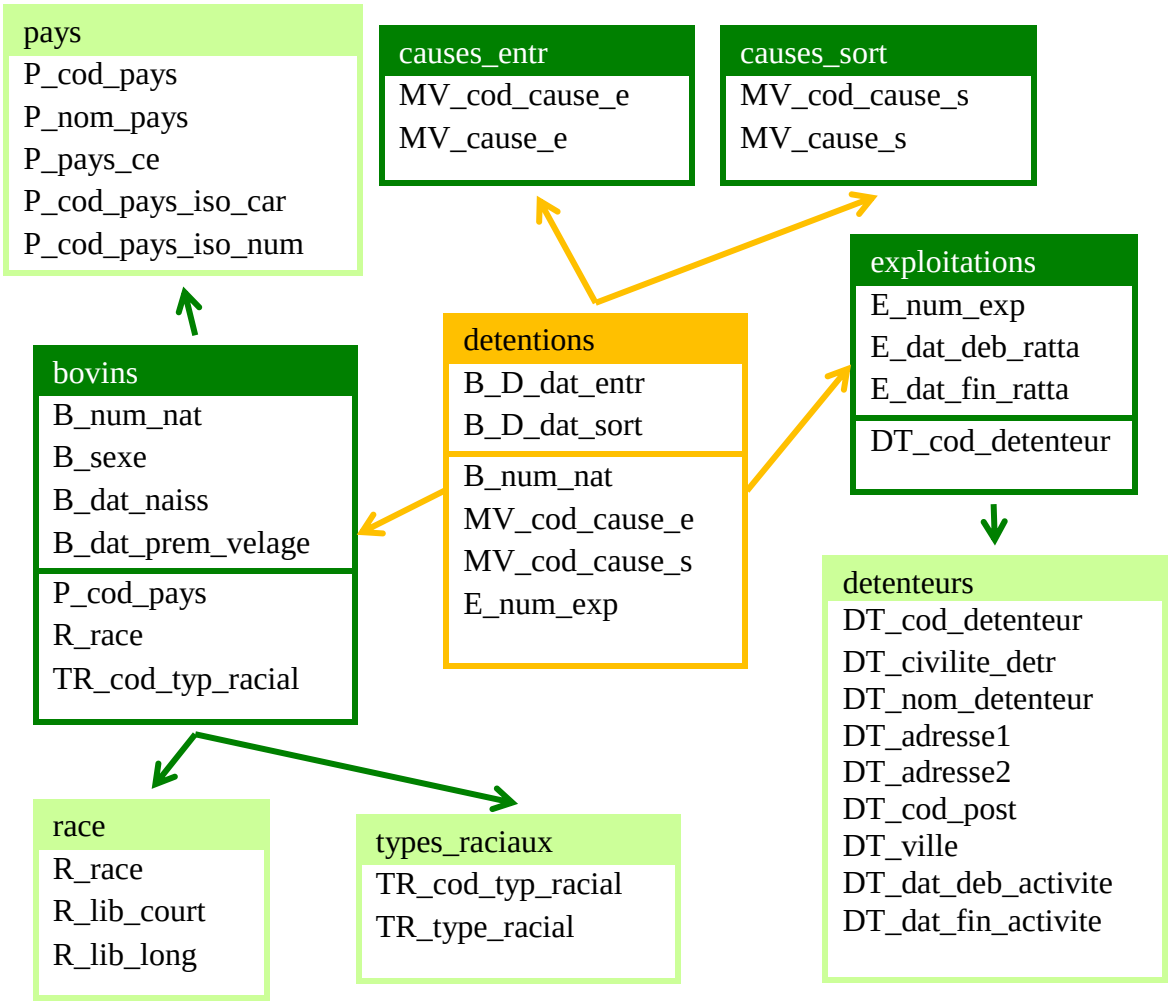


Figure 13: représentation en flocon de la BDNI

Une fois que la représentation des données est effectuée, il faut importer ces données provenant de sources différentes dans l'entrepôt via un ETL (Export-transform-load).

III-3-b- Utilisation d'un ETL pour charger les données depuis les sources

Nous avons vu que les sources de données de la BDNI (Base nationale d'identification bovine) étaient variées : certaines données proviennent de la BDNI précédente (donc d'une base de données) et d'autres sont sous forme de fichiers textes (les données de 2007 et celles de 2012). Des outils ETL (pour Extract-transform-load) sont ainsi nécessaires afin de les extraire, de les nettoyer, de les transformer et de les insérer dans un entrepôt de données [Taslimanka Sylla M., 2007] (Figure 12).

Un entrepôt de données est une base de données pouvant supporter de très grandes masses d'informations [Beranger F. et al., 2012]. Contrairement à la plupart des bases de données classiques, il est conçu pour être interrogé à des fins d'analyse [Cyran M., Lane P. & Polk J., 2005]. Les Data Warehouses sont conçus sur le modèle multidimensionnel évoqué

précédemment. Leur construction diffère donc de celle des BDD relationnelles traditionnelles puisque les BDD multidimensionnelles sont plutôt dénormalisées et respectent généralement les modèles en étoile ou en flocon [Taslimanka Sylla M., 2007].

Les ETL permettent d'extraire les données de sources hétérogènes (BDD, fichiers textes…), de les rendre cohérentes entre elles pour qu'elles puissent être utilisées ensemble et de les insérer dans l'entrepôt de données [Beranger F. et al., 2012] selon le modèle conçu (ici le modèle en flocon). Ainsi les utilisateurs disposent non seulement de données provenant de sources différentes réunies ensemble mais aussi unifiées et normalisées selon un modèle cohérent tout en évitant des redondances [Gouarné J.-M., 1997].

L'entrepôt de données est conçu dans le but de pouvoir réaliser des analyses des données, contrairement aux BDD relationnelles classiques. Ces analyses multidimensionnelles sont réalisées avec un outil appelé OLAP (On-line analytical processing). Cependant cet outil se décline en plusieurs sous-types en fonction du type d'analyses que l'on souhaite effectuer.

III-3-c- L'OLAP et les différentes analyses multidimensionnelles

L'OLAP (On-line analytical processing) est plus orienté vers l'analyse de données que les bases classiques OLTP (On-line transactional processing), en particulier lorsqu'il s'agit d'analyser plusieurs Gio d'informations [Grim Y., 2008]. L'OLAP permet d'organiser les données à analyser par axes/dimensions sous forme d'hypercube ce qui permet de résumer les résultats en fonction de ces différents axes [Taslimanka Sylla M., 2007].

On peut voir sur la figure 14 un exemple de représentation des données sous trois dimensions. Si l'indicateur calculé est le taux de mortalité, on pourrait alors voir, par exemple, le taux de mortalité des bovins laitiers dans l'est de la France en 2004 (zone verte).

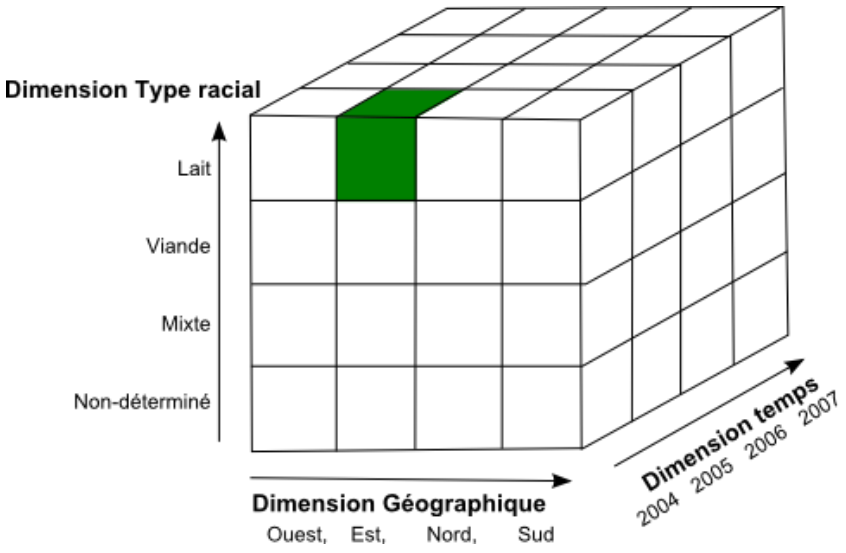


Figure 14: exemple d'hypercube à trois dimensions

Les données de l'hypercube ne sont chargées qu'une seule fois. Les requêtes ne sont adressées à la base que lorsque l'on a besoin de recharger l'hypercube, soit pour mettre à jour des données, soit pour changer de contexte [Gouarné J.-M., 1997].

L'analyse multidimensionnelle en masse permet donc d'étudier les données de la table de faits en fonction de différents axes d'analyse choisis parmi les tables de dimension [Beranger F. et al., 2012] et fournit un accès rapide à des informations stratégiques pour permettre des analyses plus poussées qu'avec l'OLTP [Grim Y., 2008].

L'OLAP permet de traiter les données de manière interactive et assez simple, et de voir les données de l'entreprise sous plusieurs angles (dimensions). Ainsi, il permet non seulement de répondre aux questions « qui » et « quoi » mais aussi aux « que se passe-t-il si » et aux « pourquoi » [Grim Y., 2008].

L'OLAP se divise en plusieurs sous-systèmes permettant chacun de faire des analyses multidimensionnelles, mais de différente manière :

- ROLAP (Relational on-line analytical processing) stocke les données multi dimensionnelles dans un format relationnel (sous forme de tables et de relations, comme dans un SGBDR classique). Le R-OLAP est particulièrement utile lorsque l'on souhait analyser de gros volumes de données dont l'accès est restreint.
- MOLAP (Multidimensional on-line analytical processing) stocke les données directement dans un format multidimensionnel, ce qui permet d'effectuer les analyses très rapidement puisque les calculs sont déjà préenregistrés. Cependant cette solution montre ses limites lorsque l'on souhaite effectuer des analyses sur de gros volumes de données car les « pré-calculs » deviennent très lourds [Grim Y., 2008].
- HOLAP (Hybrid on-line analytical processing) utilise les deux méthodes pour le stockage,
- DOLAP (Desktop on-line analytical processing) stocke les données en local pour l'analyse.
- SOLAP (Spatial on-line analytical processing) permet une analyse multidimensionnelle des données suivant un axe spatial via un affichage cartographique.

L'OLAP permet donc une analyse multidimensionnelle des données de l'entrepôt. Cet outil est particulièrement utile dans le cas de la BDNI puisqu'il permet d'analyser les données (et donc de calculer des indicateurs) sous plusieurs angles à la fois. Parmi les différents types d'OLAP, le SOLAP semble ici le plus adapté puisque l'on souhaite étudier les variations territoriales de certains indicateurs. Après avoir modélisé la BDD décisionnelle et importé les données dans un entrepôt via un ETL, il faut donc installer une plateforme décisionnelle qui puisse proposer des outils d'analyse multidimensionnelle.

III-3-d- Installation d'une plateforme décisionnelle

Une plateforme décisionnelle permet un accès centralisé et sécurisé à des données via une interface web. Elle s'appuie en général sur un entrepôt de données [Beranger F. et al., 2012] et joue simplement le rôle d'interface entre l'utilisation et les informations [Gouarné J.-M., 1997].

Certaines plateformes décisionnelles intègrent un outil ETL. Puisque les données de la BDNI proviennent de sources hétérogènes, cet outil est nécessaire et c'est donc un avantage si la plateforme en possède déjà un.

Par ailleurs la BDNI est utilisée pour le calcul d'indicateurs. La présence de l'OLAP est donc elle aussi nécessaire.

Enfin, les résultats des indicateurs ont pour vocation d'être interprétés et publiés, il faut donc que ces résultats soient faciles à interpréter visuellement. Des outils proposant des tableaux de bord prédéfinis ainsi que des outils de reporting sont donc aussi des points importants.

Puisque l'INRA ne souhaite utiliser que des outils Open Source et gratuits, j'ai fait le choix de comparer uniquement des plateformes décisionnelles Open Source. Je les ai choisies parmi les plus connues et j'ai noté la présence (+) ou l'absence (-) des différents outils mentionnés précédemment (Tableau 21).

Tableau 21: comparaison de différentes plateformes décisionnelles

	Eclipse BIRT	Jasper Community	Spago BI	Pentaho Community
ETL intégré	-	+	+	+
OLAP	-	+	+	+
Tableaux de bord	-	-	+	+
Reporting	+	+	+	+

Ainsi deux plateformes répondent à ces différents besoins : Spago BI et Pentaho. Il serait donc judicieux de choisir une des deux pour mettre en place une solution permet de calculer des indicateurs à partir de la BDNI.

Une fois l'entrepôt de données créé et la plateforme décisionnelle installée, il ne reste plus qu'à mettre en place les outils permettant d'analyser rapidement et facilement les données.

III-3-e- Développement de ressources décisionnelles

Le reporting permet de créer des rapports pré-formatés (certains avec un choix de paramètres à effectuer). Leur présentation est prédéfinie en amont par leurs concepteurs. Il reste donc à l'utilisateur à sélectionner le rapport voulu et (éventuellement) les paramètres qui l'intéressent. Il permet seulement de mieux appréhender le résultat de l'analyse. En effet, l'utilisateur final n'est pas forcément un informaticien, il peut être un chercheur ou un vétérinaire, et il aura donc plus de facilités avec ces rapports prédéfinis présentant des diagrammes et des courbes statistiques qu'en devant aller chercher de lui-même les informations dans la base [Taslimanka Sylla M., 2007].

Par exemple, sur la figure 15, l'utilisateur pourrait sélectionner l'indicateur « taux de mortalité » dont le rapport serait déjà prédéfini selon les axes « type racial », « années » et « cantons » et pourrait choisir d'entrer un paramètre supplémentaire qui serait de faire l'analyse sur les veaux de boucherie uniquement.

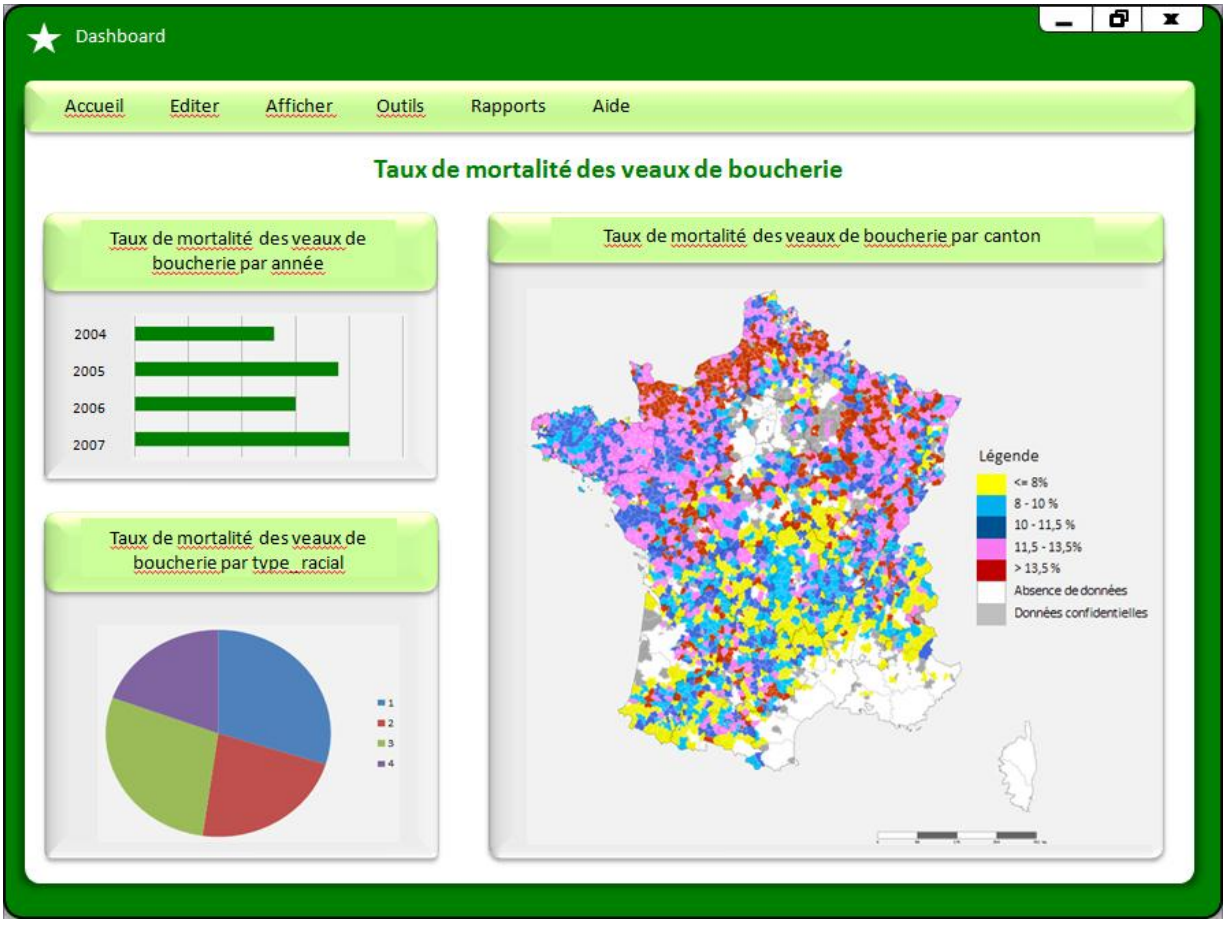


Figure 15: exemple de tableau de bord

Le tableau de bord (Figure 15) est un type de rapport particulier : tout doit tenir sur une feuille A4 ou sur un écran d'ordinateur [Beranger F. et al., 2012].

Les outils de reporting offrent donc une vision beaucoup plus claire (et assez rapide) des résultats, tout en restant très simples à manipuler puisqu'ils se présentent aux utilisateurs sous forme d'interface.

Ainsi, la Business Intelligence permet d'intégrer assez facilement des données provenant de sources diverses dans un même entrepôt de données, de les analyser rapidement sous plusieurs axes d'étude et d'obtenir les résultats sous forme synthétique avec des graphiques ou des courbes. C'est donc une solution particulièrement bien adaptée pour réponse à la problématique de cette étude, bien que ce ne soit pas celle que j'ai utilisée.

IV- Perspectives

Des difficultés ont été rencontrées au cours de ce projet, mais à chaque fois j'ai pu trouver une solution permettant de les contourner ou de faire avec. J'ai ainsi pu calculer certains indicateurs, mais je n'ai pas eu le temps de produire tous ceux qui étaient demandés. Si c'était à refaire, certains points pourraient être améliorés.

On ne le répétera jamais assez, la discussion avec les utilisateurs est un des points clefs de la réussite de tout projet. Une discussion plus approfondie avec mon maître de stage ainsi que la mise sur papier des objectifs aurait pu éviter certaines incompréhensions et m'auraient permis d'avancer plus vite dans mon travail.

Du retard a encore été accumulé à cause de la lenteur d'exécution des requêtes. Suite à cela j'ai dû mettre entre parenthèse l'intégration des données de 2012 dans la BDNI et leur traitement. Néanmoins j'ai réalisé mon code SQL de manière à ce que sa réutilisation se fasse simplement : il a été annoté, les résultats obtenus sont indiqués afin de donner un ordre de grandeur des nombres attendus et les noms des tables se terminent généralement par « _yyyy » ce qui permet de créer du code assez rapidement en remplaçant les yyyy par l'année souhaitée.

Par ailleurs nous avons vu que les données de la BDNI présentaient plusieurs incohérences. Afin d'assurer la fiabilité des indicateurs calculés par la suite, il est important de savoir d'où proviennent ces erreurs. Il faudrait donc entrer en contact avec le ministère de l'agriculture et voir avec eux l'origine des incohérences ainsi que les systèmes de repérage et de correction de ces erreurs qu'ils ont mis en place. Il serait aussi intéressant de savoir si ces sources d'incohérence proviennent par exemple d'une coopérative particulière, si elles concernent un certain type de bovin...

L'INRA a souhaité que mon travail se déroule dans la continuité de l'architecture existante et c'est ce que j'ai fait. J'ai cependant étudié d'autres solutions pouvant répondre à la problématique, notamment la business intelligence qui se révèle particulièrement adaptée aux besoins énoncés. Cependant, si cette dernière venait à être mise en place, elle nécessiterait au préalable la réalisation d'une étude du coût par rapport aux bénéfices et par la suite un accompagnement au changement. En effet, cette solution se révèle très efficace pour l'analyse multidimensionnelle de grandes masses de données et la mise en forme des résultats, mais bien que sa prise en main soit assez simple, notamment grâce à ses interfaces très ergonomiques, il faut que les utilisateurs apprennent à l'utiliser.

Conclusion

Ce projet avait pour but de calculer des indicateurs permettant d'étudier l'évolution sanitaire du cheptel bovin français et ses variations territoriales à partir des données de la BDNI.

L'INRA a souhaité que mon travail se déroule dans la continuité de l'architecture existante et c'est ce que j'ai fait. Néanmoins le système d'information actuel posait plusieurs problèmes, notamment concernant la lenteur d'exécution des requêtes pour l'analyse de masse.

J'ai donc étudié d'autres solutions pouvant répondre aux besoins de l'US-ODR, qui étaient de pouvoir stocker de grandes masses de données provenant de sources hétérogènes (BDD, fichiers textes) dans un même endroit et de les analyser rapidement selon différents axes d'étude. La solution qui s'est détachée est la business intelligence puisqu'elle répond parfaitement à ces besoins et offre des résultats assez rapide et produits sous forme de graphiques et courbes faciles à interpréter bien qu'elle soit plus longue à mettre en place initialement et qu'elle nécessite un accompagnement au changement.

Ce stage m'aura permis de mobiliser les connaissances sur les bases de données acquises en cours sur un projet concret mais aussi de développer de nouvelles compétences dans ce domaine : j'ai appris à travailler avec des bases contenant de très gros volumes de données et j'ai pu apprécier la différence d'avec des bases de moindre importance. Par ailleurs j'ai pu découvrir et m'immerger dans une utilisation des bases de données que je ne connaissais pas : la réalisation d'indicateurs. Enfin, j'ai pu m'intéresser à d'autres solutions parallèles comme la BI permettant elles-aussi de répondre aux besoins.

Bibliographie

Abello, A., & Romero, O. (2009, Avril-Juin). *A survey of multidimensional modeling methodologies*. International Journal of Data Warehousing and Mining. 23 p.

About.com. *Emerging Technologies, NoSQL: An Overview of NoSQL Databases*, In site de About.com, [En ligne] <http://newtech.about.com/od/databasemanagement/a/Nosql.htm> (Page consultée le 15/10/2013)

Annuaire des Laboratoires et des Recherches. *SAE2*, In site de l'annuaire INRA, [En ligne]. <http://annuaire.inra.fr/afficherStructure.action?code=50&type=AS> (Page consultée le 03/07/2013)

Beranger, F. et al. (2012). *Décisionnel, le meilleur des solutions open source*. Smile. 96 p.

Blain, F.-A., & Grim, Y. (2008, Mars). *Conception d'un entrepôt de données.* Developpez.com. 15 p.

Blain, F.-A., & Taslimanka Sylla, M. (2008, Avril). *Analyse et Conception d'un projet BI.* Developpez.com. 9 p.

Cyran, M., Lane, P., & Polk, J. (2005, Octobre). *b14220*. Oracle. 542 p.

Decideo.fr. Modélisation en étoile ou en flocon ?, In site de decideo, [En ligne] <www.decideo.fr/forum/Modelisation-en-etoile-ou-en-flocon_m80523.html> (Page consultée le 03/10/2013)

Downing, A. (2011, Mai). *Debunking the NoSQL Hype*. Oracle. 15 p.

Ecole Nationale Vétérinaire de Toulouse. *Présentation*, In site de l'ENVT, [En ligne]. <http://www.envt.fr/node/195> (Page consultée le 03/04/2013)

Ecole Nationale Vétérinaire de Toulouse. *Recherche*, In site de l'ENVT, [En ligne]. <http://www.envt.fr/node/548> (Page consultée le 03/04/2013)

Geniaux, G. (2006, Novembre). *Indicateurs de développement durable : un panorama des principales références bibliographiques, cadres conceptuels et initiatives internationales*. INRA. 13 p.

Gouarné, J.-M. (1997, Novembre). *Le projet décisionnel - Enjeux, modèles, architectures du Data Warehouse*. Eyrolles. 164 p.

Gray, J. (1981, Juin). *The Transaction Concept: Virtues and Limitations*. Tandem Computers Incorporated. 25 p.

Grim, Y. (2008, Mai). *OLAP, Les fondamentaux.* Developpez.com. 8 p.

Guy, Y. (2007, Septembre). *Réflexions sur les critères de choix d'indicateurs de pression phytosanitaire*. Courrier de l'environnement de l'INRA, n°54. 8 p.

IBM. *Hadoop meets SQL*, In site de IBM, [En ligne] <www.ibmbigdatahub.com/blog/hadoop-meets-sql> (Page consultée le 15/10/2013)

Inovia Conseil. *Panorama des bases de données NOSQL - 4/5*, In site de Inovia [En ligne] <http://blog.inovia-conseil.fr/?p=125> (Page consultée le 15/10/2013)

Institut National de Recherche Agronomique. *Qui sommes-nous ?*, In site de l'INRA, [En ligne]. <http://institut.inra.fr/> (Page consultée le 03/04/2013)

Larousse. *Indicateur*, [En ligne]. <http://www.larousse.fr/dictionnaires/francais/indicateur/42576/locution?q=indicateur#156691> (Page consultée le 05/10/2013)

O.D.R. INRA Plateforme Recherche. *Présentation de l'US-ODR*, In site de l'US-ODR, [En ligne] <https://esrcarto.supagro.inra.fr/intranet/carto_joomla/index.php?option=com_content&view=article&id=290&catid=46&Itemid=3147> (Page consultée le 03/04/2013)

Service public. *Annuaire de l'administration, INRA*, In site du service public, [En ligne]. <http://lannuaire.service-public.fr/services_nationaux/etablissement-public_167159.html> (Page consultée le 10/04/2013)

Service Public de Wallonie (2009*). Evolution de l'économie agricole et horticole de la Région wallonne 2008-2009*. SPW. 77 p.

Taslimanka Sylla, M. (2007, Octobre). *Initiation au décisionnel*. Developpez.com. 34 p.

TheCodingMachine. *Livres Blancs, NoSQL (not only SQL) - Brief Techno* [En ligne] <http://www.thecodingmachine.com/fr/brief-techno-nosql-not-only-sql> (Page consultée le 14/10/2013)

Tranchant, M. (2012, Avril). *Qu'est-ce que l'informatique décisionnelle ?* Developpez.com. 8 p.

Annexes

Annexe 1 – Dictionnaire des données de la bdni10

bovins

Colonne	Type	Null	Défaut	Description / Commentaire
B_IND_BOVIN	Int(10)	Non		Indice de la table
B_NUM_NAT	Char(12)	Non		Numéro national du bovin
B_SEXE	Char(1)	Oui	*NULL*	Sexe du bovin / 1 : Mâle, 2 : Femelle
B_DAT_NAISS	Date	Oui	*NULL*	Date de naissance de l'animal
B_DAT_PREM_VELAGE	Date	Oui	*NULL*	Date de premier vêlage de l'animal
B_FK_IND_PAYS	Int(11)	Oui	*NULL*	Indice pays de l'animal
B_FK_IND_PAYS NAISS	Int(11)	Oui	*NULL*	Indice pays de naissance de l'animal
B_FK_IND_EXPL_NAISS	Int(12)	Oui	*NULL*	Indice de l'exploitation de naissance de l'animal / Juste pour bdni_bis
B_NUM_EXPL_NAISS	Varchar(12)	Oui	*NULL*	N° de l'exploitation de naissance de l'animal
B_FK_IND_RACE	Tinyint(4)	Oui	*NULL*	Indice Race de l'animal
B_ FK_IND _PERE	Int(11)	Oui	*NULL*	Indice père de l'animal / Pas dans bdni_bis
B_ FK_IND_RACE_PERE	Int(12)	Oui	*NULL*	Indice race du père de l'animal / A récupérer pour bdni2
B_FK_IND_MERE	Int(11)	Oui	*NULL*	Indice mère de l'animal
B_ FK_IND_RACE_MERE	Int(12)	Oui	*NULL*	Indice race de la mère de l'animal / A récupérer pour bdni2
B_NUM_NAT_MERE	Varchar(12)	Oui	*NULL*	N° national de la mère du bovin
B_FK_IND_PAYS_MERE	Char(2)	Oui	*NULL*	Indice pays de la mère de l'animal
B_ FK_TYPE_RACIAL	Tinyint(4)	Non	*4*	Indice du type racial de l'animal / Juste pour bdni2, à calculer pour bdni_bis
B_ACTIF	Tinyint(1)	Non	*0*	Juste pour bdni2
B_IND_CATEGORIEFINALE	Smallint(5)	Non	*0*	Indice catégorie finale de l'animal / Juste pour bdni2

B_IND_CATEGORIEFINALE2	Smallint(5)	Non	*0*	Indice catégorie finale recalculer de l'animal / Juste pour bdni2

detentions

Colonne	Type	Null	Défaut	Description / Commentaire
B_D_IND	Int(10)	Non		Indice de la table
B_D_FK_IND_BOVINS	Int(11)	Non	*0*	Indice du bovin
B_D_FK_IND_EXP	Int(11)	Non		Indice de l'exploitation
B_D_DAT_ENTR	Date	Non	*0000-00-00*	Date d'entrée de l'exploitation
B_D_DAT_SORT	Date	Oui	*NULL*	Date de sortie de l'exploitation
B_D_FK_IND_CAUSE_ENTR	Tinyint(3)	Oui	*NULL*	Cause d'entrée dans l'exploitation
B_D_FK_IND_CAUSE_SORT	Tinyint(3)	Oui	*NULL*	Cause de sortie de l'exploitation

exploitations

Colonne	Type	Null	Défaut	Description / Commentaire
E_IND_EXP_DET	Int(11)	Non		Indice de la table
E_IND_EXP	Int(10)	Non		
E_NUM_EXP	Varchar(12)	Non		Numéro national de l'exploitation
E_FK_NUM_DET	Varchar(12)	Non		Numéro de la détention / Pas dans bdni2
E_DAT_DEB_RATTA	Date	Non		Date début de rattachement à l'exploitation
E_DAT_FIN_RATTA	Date	Non		Date de fin de rattachement à l'exploitation
E_EXI_PROD_BOV	Varchar(2)	Non		Existence de production bovine / Juste pour bdni_bis (=1)
E_EXI_PROD_OV	Varchar(2)	Non		Existence de production ovine / Juste pour bdni_bis
E_EXI_PROD_CAPR	Varchar(2)	Non		Existence de production caprine / Juste pour bdni_bis
E_EXI_PROD_PORC	Varchar(2)	Non		Existence de production porcine/ Juste pour bdni_bis
E_EXI_PROD_AUT	Varchar(2)	Non		Existence de production autre / Juste pour bdni_bis
E_COM	char(5)	Oui	*NULL*	Numéro de commune / Juste pour bdni2

E_CANT	char(6)	Oui	*NULL*	Numéro de canton / Juste pour bdni2
...	…	...	…	Calcul d'indicateur
ALTI	smallint(6)	Non		Calcul d'indicateur / Juste pour bdni2
STHSAU00	float	Oui	*NULL*	Calcul d'indicateur / Juste pour bdni2

detenteurs

Colonne	Type	Null	Défaut	Description / Commentaire
DT_NUM_DETR	Varchar(12)	Non		Numéro national de détenteur / Pas dans bdni2 + pas cod_pays car tous 'FR'
DT_CIVILITE_DETR	Char(6)	Oui	*NULL*	Etat civil du détenteur
DT_NOM_DETENTEUR	Varchar(36)	Oui	*NULL*	Nom du détenteur
DT_ADRESSE1	Varchar(36)	Oui	*NULL*	Adresse1 du détenteur
DT_ADRESSE2	Varchar(36)	Oui	*NULL*	Adresse2 du détenteur
DT_COD_POST	Char(7)	Oui	*NULL*	Code postal du détenteur
DT_VILLE	Varchar(36)	Oui	*NULL*	Ville du détenteur
DT_DAT_DEB_ACTIVITE	Date	Non		Date de début d'activité du détenteur
DT_DAT_FIN_ACTIVITE	Date	Oui	*NULL*	Date de fin d'activité du détenteur

abattage

Colonne	Type	Null	Défaut	Description / Commentaire
A_FK_IND_BOVINS	Int(11)	Non		Indice bovin
A_NUM_NAT	Varchar(12)	Non		N° national bovin
A_COD_PAYS	Varchar(2)	Non		Code pays
A_COD_PAYS_ABAT	Char(2)	Oui	*NULL*	Indice pays de l'abattoir
A_NUM_ABAT	Varchar(14)	Oui	*NULL*	Numéro national de l'abattoir
A_DAT_ABAT	Date	Oui	*NULL*	Date de l'abattage de l'animal
A_NUM_TUE	Varchar(14)	Oui	*NULL*	Numéro de tuerie
A_PD_CARC	Varchar(10)	Oui	*NULL*	Poids de la carcasse

exportations

Colonne	Type	Null	Défaut	Description / Commentaire
EX_IND_EXPOR	int(20)	Non		Indice de l'exportation
EX_FK_IND_BOVINS	Varchar(14)	Oui	*NULL*	Indice du bovin

EX_IND_PAYS_EXPOR	Char(2)	Oui	*NULL*	Indice pays de l'exportateur
EX_NUM_ EXPOR	Varchar(14)	Oui	*NULL*	Numéro de l'exportateur
EX_DAT_ EXPOR	Date	Oui	*NULL*	Date de l'exportation
EX_PD_ EXPOR	Varchar(10)	Oui	*NULL*	Poids de l'exportation

races

Colonne	Type	Null	Défaut	Description / Commentaire
R_IND_RACE	Tinyint(3)	Non		Indice race
R_RACE	char(2)	Non		Code race
R_LIB_COURT	varchar(20)	Non		Race court
R_LIB_LONG	varchar(50)	Non		Race long

pays

Colonne	Type	Null	Défaut	Description / Commentaire
P_IND_PAYS	Int(10)	Non		Indice pays
P_COD_PAYS	char(2)	Non		Code pays
P_NOM_PAYS	varchar(50)	Non		Nom pays
P_PAYS_CE	binary(1)	Non		Communauté européenne
P_COD_PAYS_ISO_CAR	char(3)	Non		Code pays ISO
P_COD_PAYS_ISO_NUM	char(3)	Non		Code pays ISO num

causes_mouvements

Colonne	Type	Null	Défaut	Description / Commentaire
MV_IND_CAUSE	tinyint(4)	Non		Indice de la cause du mouvement
MV_COD_CAUSE	varchar(2)	Non		Code de la cause mouvement
MV_CAUSE	varchar(50)	Non		Cause mouvement

Annexe 2 – Dictionnaire des données de 2007

EXPLOITATION:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS	char(2)	non		code du pays de l'exploitation
NUM_EXP	char(12)	non		n° de l'exploitation
TYP_EXP	numeric(2)	non		type de l'exploitation
NOM_EXP	varchar(30)	oui		nom de l'exploitation
COD_SIT	varchar(4)	oui		code de situation de l'exploitation
ADR1	varchar(30)	oui		ligne 1 de l'adresse de l'exploitation
ADR2	varchar(30)	oui		ligne 2 de l'adresse de l'exploitation
COD_POST	char(5)	non		code postal de l'exploitation
COMMUNE	varchar(30)	non		commune de l'exploitation
SIRET	char(14)	oui		SIRET de l'exploitation
COD_PAYS_DET	varchar(2)	oui		code pays du détenteur de l'exploitation
NUM_DET	varchar(12)	oui		n° du détenteur de l'exploitation
X_L_ADR	varchar(10)	oui		
Y_L_ADR	varchar(10)	oui		

DETIENT_EXP:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS_EXP	char(2)	non	FR	code du pays de l'exploitation
NUM_EXP	char(12)	non		n° de l'exploitation
COD_PAYS_DET	char(2)	non	FR	code du pays du détenteur
NUM_DET	char(12)	non		n° du détenteur
DAT_CRE_INI	char(12)	non	24/10/2001	
DAT_VSE	char(12)	non	18/10/2001	
APP	varchar(12)	non	001IPG	département de l'exploitation ?
DAT_CRE	char(12)	non	29/10/2003	
STATUT	char(1)	non	A	
DAT_DEB	char(12)	oui	01/01/1950	
CPL_DAT_DEB	numeric(2)	oui	0	

DETENTEUR:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS	char(2)	non	FR	code du pays du détenteur
NUM_DET	char(12)	non		n° du détenteur
COD_SIT	char(4)	non	M	code de la situation (situation civile et raison sociale)
NOM_DET	varchar(30)	non	BUET ROGER	nom du détenteur
ADR1	varchar(30)	oui	CHAMOUDRY	ligne 1 de l'adresse
ADR2	varchar(30)	oui		ligne 2 de l'adresse
COD_POST	char(5)	non	01400	code postal
COMMUNE	varchar(30)	non	CLEMENCIAT	commune
SIREN	char(9)	oui	417960119	n° SIREN
COD_PAYS_RES	char(2)	non	FR	code du pays de résidence du détenteur ?
DAT_VSE	char(12)	non	13/09/1999	
APP	varchar(12)	non	001IPG	département de l'exploitation ?
DAT_CRE	char(12)	non	06/02/2007	
STATUT	char(1)	non	R	

ACTIVITE_EXP:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS	char(2)	non	FR	code du pays de l'exploitation
NUM_EXP	char(12)	non		n° de l'exploitation
DAT_VSE	char(12)	non	18/10/2001	
DAT_DEB_ACT	char(12)	non	01/01/1950	date du début de l'activité de l'exploitation
CPL_DAT_DEB	numeric(2)	non	0	
DAT_FIN_ACT	char(12)	oui		date de fin d'activité de l'exploitation
CPL_DAT_FIN	numeric(2)	oui		
APP	varchar(12)	non	001IPG	département de l'exploitation ?
DAT_CRE	char(12)	non	29/10/2003	
STATUT	char(1)	non	A	
TYP_PROD	varchar(1)	oui	B	type de production de l'exploitation

RATTACH_EGET:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS_EGET	char(2)	non	FR	code du pays importateur ?

NUM_EGET	char(12)	non	01BDZ	n° du pays importateur ?
TYP_PROD	char(1)	non	P	type de production
COD_PAYS_EXP	char(2)	non	FR	code du pays exploitation
NUM_EXP	char(12)	non		n° de l'exploitation
DAT_DEB	char(12)	non	01/01/1950	date du début de l'activité de l'exploitation
CPL_DAT_DEB	numeric(2)	non	0	
DAT_VSE	char(12)	non	27/05/2005	
APP	varchar(12)	non	001IPG	département de l'exploitation ?
DAT_CRE	char(12)	non	10/03/2006	
STATUT	char(1)	non	A	

ENTITE_EGET:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS_EGET	char(2)	non	FR	code du pays importateur ?
NUM_EGET	char(12)	non	01AAZ	n° du pays importateur ?
TYP_PROD	char(1)	non	P	type de production de l'exploitation importatrice
ADR1	varchar(30)	oui	BARBET	ligne 1 de l'adresse de l'exploitation importatrice
ADR2	varchar(30)	oui		ligne 2 de l'adresse de l'exploitation importatrice
COD_POST	char(5)	non	01320	code postal de l'exploitation importatrice
COMMUNE	varchar(30)	non	SAINT NIZIER	commune de l'exploitation importatrice
DAT_VSE	char(12)	non	22/03/2006	
APP	varchar(12)	non	001IPG	département de l'exploitation ?
DAT_CRE	char(12)	non	22/03/2006	
STATUT	char(1)	non	A	

Num_Bov_Zone:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS_ANI	char(2)	non	FR	code du pays du détenteur de l'animal
NUM_NAT	char(12)	non	0102300967	n° national de l'animal

BOVIN:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS	char(2)	oui	FR	code du pays du détenteur de l'animal

NUM_NAT	char(12)	oui	8531149422	n° national de l'animal
NOM	varchar(10)	oui		nom de l'animal
TYP_RACE	numeric(2)	oui	38	race de l'animal
Type_Prod	varchar(22)	oui	Viande	type de production de l'animal
champ_inconnu1	char(3)	oui	Oui	
SEXE	numeric(2)	oui	1	sexe de l'animal
COD_ROBE	varchar(3)	oui		robe de l'animal
TEMOIN_NAISS	numeric(2)	oui	1	
DAT_NAISS	char(12)	oui	19/03/2007	date de naissance de l'animal
CPL_DAT_NAISS	numeric(2)	oui	0	
COD_PAYS_NAISS	varchar(2)	oui	FR	pays de naissance de l'animal
NUM_EXP_NAISS	varchar(12)	oui	85084047	n° de l'exploitation de naissance de l'animal
COD_PAYS_PERE	varchar(2)	oui		code du pays du père de l'animal
NUM_NAT_PERE	varchar(12)	oui		n° national du père de l'animal
TYP_RACE_PERE	numeric(2)	oui	38	race du père de l'animal
Type_Prod_Père	char(1)	oui	V	type de production du père de l'animal
champ_inconnu2	char(3)	oui	Oui	
COD_PAYS_MERE	varchar(2)	oui	FR	code du pays de la mère de l'animal
NUM_NAT_MERE	varchar(12)	oui	8540240239	n° national de la mère de l'animal
TYP_RACE_MERE	numeric(2)	oui	38	race de la mère de l'animal
Type_Prod_Mere	char(1)	oui	V	type de production de la mère de l'animal
champ_inconnu3	char(3)	oui	Oui	
NUM_TRA	varchar(4)	oui	9422	n° de travail de l'animal ?
FILIE	numeric(2)	oui		
DAT_MORT	char(12)	oui		date de mort de l'animal
CPL_DAT_MORT	numeric(2)	oui		
DAT_ABAT	char(12)	oui	22/07/2008	date de l'abattage de l'animal
CPL_DAT_ABAT	numeric(2)	oui	0	
DAT_EQUAR	char(12)	oui		date d'équarissage de l'animal
CPL_DAT_EQUAR	numeric(2)	oui		
DAT_FIN_GES	char(12)	oui		date de fin de gestation de l'animal
CPL_DAT_FIN_GES	numeric(2)	oui		

STA_DOSS	numeric(2)	oui		
DAT_VSE	char(12)	oui	28/03/2007	
APP	varchar(12)	oui	031BDN	département de l'exploitation
DAT_CRE	char(12)	oui	03/04/2010	
STATUT	char(1)	oui	A	
CTRL_PASS	numeric(2)	oui	1	
DAT_RECEP	char(12)	oui	28/03/2007	
DAT_PREM_VELAGE	char(12)	oui		date du premier vêlage de l'animal
CPL_DAT_PREM_VELAGE	numeric(2)	oui		
DAT_FIN_VIE	char(12)	oui	22/07/2008	Date d'arrêt de la production de l'animal ?
CPL_DAT_FIN_VIE	numeric(2)	oui	0	
STA_ABAT	char(1)	oui	1	

MVT_ENTR:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS_ANI	char(2)	non	FR	code du pays exportateur de l'animal
NUM_NAT	char(12)	non	0334135696	n° national de l'animal
COD_PAYS_ENTR	char(2)	non	FR	code du pays importateur
NUM_EXP_ENTR	char(12)	non		n° de l'exploitation importatrice
DAT_VSE	char(12)	non	02/03/2005	
CAUSE_MVT	char(1)	non	N	code de la cause du mouvement de l'animal
TYP_MVT	char(1)	non	exemple	
DAT_ENTR	char(12)	non	19/08/2007	date d'entrée de l'animal sur l'exploitation
CPL_DAT_ENTR	numeric(2)	non	0	
ID_EXP_PROV	varchar(14)	oui		id de l'exploitation de provenance de l'animal
NOM_EXP_PROV	varchar(60)	oui		nom de l'exploitation de provenance de l'animal
APP	varchar(12)	non	003IPG	département de l'exploitation
DAT_CRE	char(12)	non	27/08/2007	
STATUT	char(1)	non	A	
DAT_RECEP	char(12)	oui		

MVT_SORT:

Colonne	Type	null	Exemple	Commentaire

COD_PAYS_ANI	char(2)	non	FR	code du pays importateur de l'animal
NUM_NAT	char(12)	non	0200435824	n° national de l'animal
COD_PAYS_SORT	char(2)	non	FR	code du pays exportateur
NUM_EXP_SORT	char(12)	non		n° de l'exploitation exportatrice
DAT_VSE	char(12)	non	23/01/2007	
CAUSE_MVT	char(1)	non	M	code de la cause du mouvement de l'animal
TYP_MVT	char(1)	non	S	
DAT_SORT	char(12)	non	16/01/2007	date de sortie de l'animal de l'exploitation
CPL_DAT_SORT	numeric(2)	non	0	
ID_EXP_DEST	varchar(14)	oui		id de l'exploitation de destination de l'animal
NOM_EXP_DEST	varchar(60)	oui	INCONNU	nom de l'exploitation de destination de l'animal
APP	varchar(12)	non	009IPG	département de l'exploitation
DAT_CRE	char(12)	non	23/01/2007	
STATUT	char(1)	non	A	
DAT_RECEP	char(12)	oui		

MGA_PERI_DET:

Colonne	**Type**	**null**	**Exemple**	**Commentaire**
COD_PAYS_DET	varchar(2)	oui	FR	code du pays du détenteur de l'animal
NUM_DET	varchar(12)	oui	01500015030	n° du pays du détenteur de l'animal
COD_PAYS_EXP	char(2)	non	FR	code du pays exportateur
NUM_EXP	char(12)	non		n° de l'exploitation
COD_PAYS_ANI	char(2)	non	FR	code du pays du détenteur de l'animal
NUM_NAT	char(12)	non	1526142860	n° national de l'animal
DAT_ENTR	char(12)	oui	06/02/2007	date d'entrée de l'animal sur l'exploitation
DAT_SORT	char(12)	oui	05/03/2007	date de sortie de l'animal de l'exploitation
KLITE_ENTR	smallint(2)	oui	0	
KLITE_SORT	smallint(2)	oui	0	
CAUSE_ENTR	char(1)	oui	A	code de la cause d'entrée de l'animal sur l'exploitation
CAUSE_SORT	char(1)	oui	E	code de la cause de sortie de l'animal de l'exploitation

DEP_ORIGINE	char(3)	oui	015	
DEP_SORT	char(3)	oui	012	
DEP_EXP	char(3)	non	015	
DEP_DET	char(3)	oui	015	
DAT_RECEP_ENTR	char(12)	oui		
DAT_RECEP_SORT	char(12)	oui		
PAYS_ORIGINE	char(3)	oui		pays d'origine de l'animal

ABATTAGE:

Colonne	**Type**	**null**	**Exemple**	**Commentaire**
COD_PAYS_ANI	char(2)	non	FR	code du pays du détenteur de l'animal
NUM_NAT	char(12)	non	0102300968	n° national de l'animal
COD_PAYS_EXP_PROV	varchar(2)	oui	FR	code du pays exportateur
NUM_EXP_PROV	varchar(12)	oui		n° de l'exploitation exportatrice
DEP_PROV	varchar(5)	oui	FR071	département de provenance ?
DAT_ABAT	char(12)	non	10/02/2012	date de l'abattage de l'animal
Année_Abat	int(4)	non	2012	année de l'abattage de l'animal
Mois_Abat	int(4)	non	2	mois de l'abattage de l'animal
Semaine_Abat	int(4)	non	6	semaine de l'abattage de l'animal
POIDS	numeric(4)	oui	382.20	poids de l'animal après abattage
Num_EDE_Abat	char(12)	oui		n° EDE de l'abattoir ?
TYP_RACE	numeric(2)	oui	38	race de l'animal
SEXE	numeric(2)	oui	2	sexe de l'animal
Cat_Animal	varchar(20)	non	Vache	catégorie d'animal
DAT_NAISS	char(12)	oui	15/05/2007	date de naissance de l'animal
TYP_RACE_PERE	numeric(2)	oui	38	race du père de l'animal
TYP_RACE_MERE	numeric(2)	oui	38	race de la mère de l'animal
DAT_PREM_VELAGE	char(12)	oui	04/02/2010	date du premier vêlage de l'animal
Age_Abat_Jours	int(4)	oui	1732	âge de l'animal (en jours) au moment de l'abattage
Cat_Age_Abat_Esst	varchar(20)	non	49-72Mois	catégorie d'âge pour l'abattage ESST ?
Cat_Age_AbaTTech	varchar(20)	non	48-119Mois	catégorie d'âge pour l'abattage TECH ?
Type_Prod	varchar(22)	oui	Viande	type de production

EQUARRISSAGE:

Colonne	Type	null	Exemple	Commentaire
COD_PAYS_ANI	char(2)	non	Fr	code du pays du détenteur de l'animal
NUM_NAT	char(12)	non	0103303373	n° national de l'animal
TEMOIN_NUM_NAT	numeric(2)	non	0	
COD_PAYS_EXP_PROV	varchar(2)	oui	FR	code du pays exportateur
NUM_EXP_PROV	varchar(12)	oui		n° de l'exploitation exportatrice
COD_INSEE_PROV	varchar(5)	oui		code INSEE de la provenance
DEP_PROV	varchar(5)	oui	FR001	département de provenance ?
DAT_ENLV	char(12)	non	04/05/2007	date d'enlèvement de l'animal
Année_Equa	int(4)	non	2007	année de l'équarrissage de l'animal
Mois_Equa	int(4)	non	5	mois de l'équarrissage de l'animal
Semaine_Equa	int(4)	non	18	semaine de l'équarrissage de l'animal
CPL_DAT_ENLV	numeric(2)	non	0	
STA_BOUCLE	numeric(2)	non	1	
STA_PASS	numeric(2)	non	0	
COD_PAYS_EQUA	char(2)	non	FR	code pays de l'établissement d'équarrissage
NUM_EQUA	char(12)	non		n° de l'établissement d'équarrissage
NUM_COLLECTE	char(12)	non		n° de collecte de l'équarrissage
DAT_VSE	char(12)	non	02/08/2007	
APP	varchar(12)	non	01451383	
DAT_CRE	char(12)	non	08/08/2007	
STATUT	char(1)	non	A	
TYP_RACE	numeric(2)	oui	38	race de l'animal
SEXE	numeric(2)	oui	1	sexe de l'animal
Cat_Animal	varchar(20)	non	Mâle	catégorie d'animal
DAT_NAISS	char(12)	oui	25/02/2007	date de naissance de l'animal
TYP_RACE_PERE	numeric(2)	oui	38	race du père de l'animal
TYP_RACE_MERE	numeric(2)	oui	38	race de la mère de l'animal
DAT_PREM_VELAGE	char(12)	oui		date du premier vêlage de l'animal

Age_EQUA_Jours	int(4)	oui	68	âge de l'animal (en jours) au moment de l'abattage
Cat_Age_EQUA_Esst	varchar(20)	non	00-11Mois	catégorie d'âge de l'animal lors de l'équarrissage ESST ?
Cat_Age_EQUA_Tech	varchar(20)	non	00-07Mois	catégorie d'âge de l'animal lors de l'équarrissage TECH ?
Type_Prod	varchar(22)	oui	Viande	type de production

Résumé

L'unité de service de l'Observatoire du développement rural (US-ODR) de l'Institut national de recherche agronomique (INRA) de Toulouse effectue des recherches pour décrire le monde économique rural. Pour cela, l'US-ODR utilise les informations de la Base de données nationale d'identification bovine (BDNI) en vue de produire des indicateurs.

Les objectifs de ce stage étaient de mettre à jour la BDNI et de construire des indicateurs par année. A la demande de l'INRA, mon projet s'est déroulé dans la continuité de l'architecture existante. J'ai alors été confrontée à plusieurs difficultés : certaines données étaient absentes de la BDNI, d'autre étaient incohérentes, l'exécution des requêtes était très lente et l'architecture de la base existante ne se prêtait pas au calcul d'indicateurs par année.

J'ai donc étudié des solutions alternatives pour optimiser le système et produire les indicateurs plus rapidement et plus facilement.

A l'issue de ce projet, j'ai atteint les objectifs initiaux et j'ai proposé une piste qui permettrait d'optimiser le système en vue de produire des indicateurs. Cependant, cette solution reste à concrétiser.

Mots-clés : Base de données, SGBD, BDNI, Indicateur, Informatique décisionnelle

Abstract

The unit Observatory of rural development (US-ODR) of the National institute of agronomic research (INRA) in Toulouse conducts research to describe the rural economic world. To this end, US-ODR uses the National database of cattle identification (BDNI) in order to establish indicators.

The goal of this project was to update the BDNI and to establish annual indicators. At the request of the INRA, my project took place in the continuity of the existing architecture. I then faced several problems: a part of the data was missing in the BDNI, another part was inconsistent, the query execution was very slow and the architecture of the existing database didn't lend itself to the calculation of the annual indicators.

That is why I studied alternative solutions to optimize the system and produce indicators quickly and easily.

As result, I reach the initial goals and I proposed a track that would optimize the system to produce indicators. However, this solution is still to shape.

Keywords: Database, Database management system, BDNI, Indicators, Business intelligence